Econometrics

Preliminaries

Probability

Probability Space

Definition: (Probability Space) A Probability Space is defined as (Ω, F, P), where Ω is the sample space, F is the sigma algebra defined on Ω, and P is the probability measure.

Claim:(Properties of Probability) We have $P(\phi) = 0$, $P(A) \in [0, 1]$, and $P(A^c) = 1 - P(A)$.

Definition: (Disjoint) Two events are Disjoint if $P(A \cap B) = 0$.

Definition: (Independent) Two events are Independent if $P(A \cap B) = P(A)P(B)$.

Definition: (Conditional Probability) the Conditional Probability is defined as $P(A|B) = P(A \cap B)/P(B)$.

Claim: (Properties of Conditional Probability) We have $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$.

Claim: (Total Probability Formula) We have $P(A) = \sum_i P(A \cap B_i)$ where $\{B_i\}$ is a partition of Ω.

Claim: (Bayes Rule) We have $P(B|A) = \frac{P(A|B)}{P(A)}P(B)$.

Random Variable

Definition: (Random Variable) Random Variable is a function $X : \Omega \to \mathbb{R}$.

Definition: (Cumulative Distribution Function) The Cumulative Distribution Function is the function such that $F_X(a) = P(X \leq a)$.

Claim:(Properties of CDF) A CDF of a random variable is non-decreasing, between 0 and 1, and continuous from the right. Plus we have $lim_{a \to -\infty} F_X(a) = 0$, and $lim_{a \to +\infty} F_X(a) = 1$.

Definition: (Probability Density Function) For a continuous random variable, the Probability Density Function is defined as $f_X(a) = \frac{d}{da}F_X(a)$.

Definition: (Joint CDF) The Joint CDF is the function such that $F(x_1, x_2, \dots, x_k) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_k \leq x_k)$.

Definition: (Joint PDF) For a bunch of continuous variables, the Joint PDF is defined as $f(x_1, \dots, x_k) = \frac{d}{dx_1} \cdots \frac{d}{dx_k} F(x_1, \dots, x_k)$.

Definition: (Conditional PDF) Given two vectors of continuous random variables, the Conditional PDF is defined as $f(y|x) = f(x, y)/f(x)$.

Claim: (Transformation) If $Y = G(X)$, then $F_Y(a) = P(Y \leq a) = P(G(X) \leq a)$. Furthermore, if X, Y are two vector, if there exists a function such that $X = H(Y)$, then $f_Y(y) = |J(Y)| f_X(H(y))$, where $J(Y) = [\frac{\partial}{\partial y_j} H_i(y)]$ is the Jacobian matrix of $H(.)$.

Claim: (Monotonic Transformation) Suppose $Y = G(X)$, then $f_Y(y) = |\frac{d}{dy} g^{-1}(y)| f_X(G^{-1}(y))$.

Definition: (Moments) The r-th order Moments of a random variable is defined as $E[X^r] = \int_{-\infty}^{+\infty} X^r dF_X(X)$.

Definition: (Expectation, Variance, Covariance) The Expectation of a random variable is defined as $E[X] = \int_{-\infty}^{+\infty} X dF_X(X)$. the Variance is defined as $Var(X) = E[X^2] - E[X]^2 = E[(X - E[X])^2]$. The Covariance of two random variables is defined as $Cov(X, Y) = E[XY] - E[X]E[Y] = E[X - E[X]]E[Y - E[Y]]$.

Claim: (Law of Iterated Expectation) We have $E[E[Y|X]] = E[Y]$, and $E[[Y|X_1, X_2]|X_1] = E[Y|X_1]$.

Definition: (Hazard Function) The Hazard Function is defined as $H(x_0) = f_X(x_0)/(1 - F_X(x_0))$.

Inequalities

Claim: (Chebeshev's Inequality) $P(g(X) \geq r) \leq E[g(x)]/r$.

Claim: (Jensen's Inequality) If $g(.)$ is convex, then $E[g(X)] \geq g(E[X])$.

Claim: (Holder's Inequality) If $\frac{1}{p} + \frac{1}{q} = 1$, we have

1. $ab \leq \frac{1}{p} a^p + \frac{1}{q} b^q$
2. $E[|XY|] \leq E[|X|^p]^{\frac{1}{p}} E[|Y|^q]^{\frac{1}{q}}$

Claim: (Minkovski's Inequality) $E[|X + Y|^p]^{\frac{1}{p}} \leq E[|X|^p]^{\frac{1}{p}} + E[|Y|^p]^{\frac{1}{p}}$.

Linear Projection

Definition: (Linear Projection) The best linear predictor is defined as $P(y|x) = x'\beta$, where beta is defined as

$$\beta = E[xx']^{-1} E[xy] = argmin E[(y - x'b)^2] \tag{1}$$

Claim: (Law of Iterated Projection) The following statements are true:

1. $P(ay_1 + by_2|x) = aP(y_1|x) + bP(y_2|x)$
2. $P(P(y|x)) = p(y)$ and $P(P(y|x_1, x_2)|x_1) = P(y|x_1)$.

Distribution

Discrete Random Variable

Distribution	PDF	MGF	Expectation	Variance
Bernoulli	$f(x) = p^x(1-p)^{1-x}, x = 0, 1$	$M(t) = 1 - p + pe^t, t \in \mathbb{R}$	p	$p(1-p)$
Binomial	$f(x) = \frac{n! p^x (1-p)^{n-x}}{x!(n-x)!}, x = 0, 1, \ldots, n$	$M(t) = (1 - p + pe^t)^n, t \in \mathbb{R}$	np	$np(1-p)$
Geometric	$f(x) = (1-p)^{x-1} p, x = 1, 2, 3, \ldots$	$M(t) = \frac{pe^t}{(1-(1-p)e^t)}, t < -ln(1-p)$	$1/p$	$\frac{(1-p)}{p^2}$
Hypergeometric	$f(x) = \binom{N_1}{x}\binom{N_2}{n-x} / \binom{N_1+N_2}{n}$	-	$n\frac{N_1}{N_1+N_2}$	$n\frac{N_1}{N_1+N_2}\frac{N_2}{N_1+N_2}\frac{N_1+N_2-n}{N_1+N_2-1}$
Negative Binomial	$f(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, x = r, r+1, \ldots$	$M(t) = (pe^t)^r/[1-(1-pe^t)]^r, t < -ln(1-p)$	r/p	$r(1-p)/p^2$
Poisson	$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0, 1, 2, \ldots$	$M(t) = exp(\lambda(e^t - 1)), t \in \mathbb{R}$	λ	λ
Uniform	$f(x) = 1/m, x = 1, 2, 3, \ldots, m$	-	$(m+1)/2$	$(m^2-1)/12$

Continuous Random Variable

Distribution	PDF	MGF	Expectation	Variance
Uniform	$f(x) = \frac{1}{b-a}, x \in [a, b]$	$M(t) = \frac{e^{tb}-e^{ta}}{t(b-a)}, t \neq 0$	$\frac{a+b}{2}$	$(b-a)^2/12$
Gamma	$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, x > 0$	$M(t) = \frac{1}{(1-\beta t)^\alpha}, t < 1/\beta$	$\alpha\beta$	$\alpha\beta^2$
Exponential	$f(x) = e^{-x/\lambda}/\lambda, x \geq 0$	$M(t) = \frac{1}{1-\lambda t}, t \leq 1/\lambda$	λ	λ^2
Chi-Squared	$f(x) = \frac{1}{\Gamma(r/2)2^{r/2}} x^{r/2-1} e^{x/2}, x > 0$	$M(t) = 1/(1-2t)^{r/2}, t < 1/2$	r	$2r$
Beta	$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, x \in (0, 1)$	-	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(1+\alpha+\beta)(\alpha+\beta)^2}$
Normal	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}$	$M(t) = exp(\mu t + \frac{\sigma^2 t^2}{2}), t \in \mathbb{R}$	μ	σ^2
T	$f(x) = \frac{\Gamma(\frac{r+1}{2})}{(\sqrt{r\pi}\Gamma(r/2))}(1+x^2/r)^{-\frac{r+1}{2}}, x \in \mathbb{R}$	-	0	$\frac{r}{r-2}$
F	$f(x) = (\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1+d_2}})^{\frac{1}{2}}/(xB(d_1/2, d_2/2)), x \in \mathbb{R}$	-	$d_2/(d_2-2)$	$\frac{2d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$
Multinormal	$f(x) = (2\pi)^{-k/2}\lvert\Sigma\rvert^{-1/2} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}, x \in \mathbb{R}^k$	$M(t) = exp(\mu^T t) + \frac{1}{2}t^T\Sigma t$	μ	Σ

Definition: (Gamma Function) Gamma Function is $\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1}e^{-t}dt$. We have $\Gamma(\alpha) > 0$, $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$, $\Gamma(n) = n!$ and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Statistics and Convergence Theory

Random Sampling

Definition: (Random Sample) Suppose $\{X\}$ is the set of population, a subset $\{X_n\} \in \{X\}$ is called a Random Sample, where $X_i \sim f_X$ are mutually independent and have identical distribution.

Note: The joint PMF or PDF of $\{X_n\}$ is $f_{\{X_n\}} = \prod_{i=1}^n f_X(x_i)$.

Definition: (Estimator) An estimator is a function of the sample, i.e. $\hat{\theta} = T(\{X_n\})$.

Definition: (Sampling Distribution) The distribution of $\hat{\theta}$ is called a sampling distribution.

Convergence

Definition: (Convergence in Probability) A sequence of Random Variables is said to converge in probability to $\mu \in \mathbb{R}$ if $lim_{n\to+\infty} P(|X_n - \mu| < \epsilon) = 1$ for $\epsilon > 0$.

Definition: (Orders in Probability) We write $X_n = O(n^r)$ if X_n/n^r is bounded in probability, i.e. for any $\epsilon > 0$, there exists $b \in \mathbb{R}$ and $N \in \mathbb{R}$ for $P(|X_n/n^r| > b) < \epsilon$.

Definition: (Higher Orders in Probability) We write $X_n = o(n^r)$ if $lim_P X_n/n^r = 0$.

Definition: (Convergence in Distribution) We say X_n converges in distribution to X when the CDF of X_n converges to X, i.e. $lim_{n\to+\infty} F_{X_n}(x) = F_X(x)$ for all x.

Claim: (Continuous Mapping Theorem) Suppose $lim_p X_n = \mu_X$, $lim_p Y_n = \mu_Y$, and $lim_d Z_n = Z$, the following statements are true:

1. $lim_p aX_n = a\mu_X$, where a is a scaler
2. $lim_p X_n + Y_n = \mu_X + \mu_Y$, $lim_p X_n Y_n = \mu_X \mu_Y$, and $lim_p X_n/Y_n = \mu_X/\mu_Y$ if $\mu_Y \neq 0$
3. If $g(.)$ is a continuous function, then $lim_p g(X_n, Y_n) = g(\mu_X, \mu_Y)$
4. $lim_d aZ_n = aZ$, where a is a scaler
5. $lim_d X_n + Y_n Z_n = \mu_X + \mu_Y Z$
6. If $g(.)$ is a continuous function, then $lim_d g(Z_n) = g(Z)$
7. If $lim_p X_n = Z_n$ and $lim_d Z_n = Z$, then $lim_d X_n = Z$

Law of Large Number

Claim: (Weak Law of Large number) Assume that $\{X_i\}_{i=1}^N$ are $i.i.d.$ with $E[X_i] = \mu < +\infty$, and $Var(X_i) < +\infty$, then we have:

$$lim_p \frac{1}{N}\sum_{i=1}^{N} X_i = \mu \tag{2}$$

Central Limit Theorem

Claim: (Central Limit Theorem) Assume that $\{X_i\}_{i=1}^N$ are $i.i.d.$ with $E[X_i] = \mu < +\infty$, and $Var(X_i) = \Sigma < +\infty$, then we have:

$$\sqrt{n}(\bar{X}_n - \mu) = \sqrt{n}(\frac{1}{N}\sum_{i=1}^{N} X_i - \mu) = \rightarrow^d N(0, \Sigma) \tag{3}$$

Delta Method

Claim: (Delta Method) Suppose $g(.)$ is twice continuously differentiable at μ, such that $lim_p X_n = \mu$ and $\sqrt{n}(x_n - \mu) \rightarrow^d N(0, \Sigma)$, then:

$$\sqrt{n}(g(X_n) - g(\mu)) \rightarrow^d Dg(\mu)N(0, \Sigma) = N(0, Dg(\mu)'\Sigma Dg(\mu)) \tag{4}$$

Point Estimation and Confidence Intervals

Maximum Likelihood

Definition: (Likelihood Function) Likelihood Function of a sample is defined as:

$$L_n(\theta) = \prod_{i=1}^{n} f(X_i, \theta) \tag{5}$$

Definition: (Maximum Likelihood Estimator) Maximum Likelihood Estimator of a sample is defined as:

$$\hat{\theta} = argmax[lnL_n(\theta)] = argmax[\sum_{i=1}^{n} lnf(X_i, \theta)] \tag{6}$$

Method of Moments

Definition: (Method of Moments Estimator) When the population random variable X have the following property:

$$E[m(X, \theta)] = 0 \tag{7}$$

Then Method of Moments Estimator of a sample is the solution to the following equation:

$$\sum_{i=1}^{n} m(X_n, \hat{\theta})/n = 0 \tag{8}$$

Comparison of Estimators

Definition: (Unbiasedness) If $E[\hat{\theta}] = \theta$, then we say the estimator is unbiased.

Definition: (Mean Square Error) The mean square error of the estimation is defined by $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = (E[\hat{\theta}] - E[\theta])^2 + var(\hat{\theta})$

Definition: (Efficiency) Given two estimator $\hat{\theta}_1$ and $\hat{\theta}_2$, for a given sample size, if $Var(\hat{\theta}_1) \leq Var(\hat{\theta}_2)$, we say $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$.

Definition: (Consistency) The estimator is consistent if $lim_p \hat{\theta}_n = \theta$.

Confidence Intervals

Definition: (Confidence Interval) Given the data $\{S_n\}$ we observe, suppose $S_i \sim f(\theta)$. Let L and U be two statistics. We say (L, U) is a $1-\alpha$ Confidence Interval for θ if $P(\theta \in (L, U)) = 1-\alpha$.

Statistical Inferences

Hypothesis Test

Definition: (Null Hypothesis) Suppose $\theta \in \Theta$ is a random parameter, Null Hypothesis is $H_0 : \theta \in \Theta_0$.

Definition: (Alternative Hypothesis) Suppose $\theta \in \Theta$ is a random parameter, Alternative Hypothesis is $H_1 : \theta \notin \Theta_0$.

Definition: (Type I Error) Type I Error is when you reject H_0 when it is correct.

Definition: (Type II Error) Type II Error is when you accept H_0 when it is not correct.

Note: Type I Error is much worse than Type II Error.

Definition: (Decision Rule) Given the data $\{S_n\}$ we observe, we setup a rejection region C, such that if $S_n \in C$ we reject H_0, if $S_n \notin C$ we refuse to reject H_0.

Definition: (Size) The size of a Hypothesis Test is the probability of making type I error, i.e. $size = P(S_n \in C|\theta_0)$.

Definition: (P-Value) Suppose H_0 is true and a given rejection region C, P-Value is defined as $P(C) = P(S_n \in C|\theta_0)$

Definition: (Power) The size of a Hypothesis Test is the probability of not making type II error, also known as the probability of rejecting a given alternative hypothesis $\theta \in \Theta \backslash \Theta_0$, i.e. $power(\theta) = P(S_n \in C|\theta \in \Theta \backslash \Theta_0)$.

Note: We would want the power to be high and the size to be low.

Comparison of Decision Rules

Definition: (Unbiased Test) A test is called unbiased if it is more likely to reject under Alternative Hypothesis than under then Null Hypothesis.

Definition: (Consistent Test) A test is called consistent if $lim_{n \to +\infty} P(S_n \in C|\theta \in \Theta \backslash \Theta_0) = 1$.

Ordinary Least Squares Estimation

Regression Model

General Regression Model

Definition: (General Regression Model) A regression model is defined as:

$$y = m(x) + \epsilon \tag{9}$$

with $E[\epsilon|x] = 0$ and $E[\epsilon^2|x] = \sigma^2(x)$.

Linear Regression Model

Definition: (Linear Regression Model) A linear regression model is defined as:

$$y = x'\beta + \epsilon \tag{10}$$

with $E[\epsilon|x] = 0$ and $E[\epsilon^2|x] = \sigma^2(x)$.

Definition: (Sample) A sample $\{(X, Y)\}$ is drawn from the population $\{(x, y)\}$.

Definition: (Sample Regression Model) A linear regression model of the sample is defined as:

$$y = X\beta + e \tag{11}$$

with $E[e|X] = 0$ and $E[e^2|X] = \sigma^2(X)$.

Definition: (Least Square Estimator) A least square estimator is defined as:

$$\hat{\beta} = argmin(\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'b)^2) = argmin(\frac{1}{n}(y - Xb)'(y - Xb)) \tag{12}$$

Assumption

Assumption 1: (Random sampling) Each Sample is drawn with i.i.d.

Assumption 2: (No Perfectly Collinearity) $X'X$ is invertible.

Assumption 3': (Zero Correlation) $E[Xe] = 0$.

Assumption 3: (Zero Conditional Mean) $E[e|X] = 0$.

Note: Zero Conditional Mean is stronger than Zero Correlation.

Assumption 4': (Heteroskedasticity) $E[e^2|X] = \sigma^2(X)$.

Assumption 4: (Homoscedasticity) $E[e^2|X] = \sigma^2$.

Assumption 5: (Gaussian Error) $e|X \sim N(0, \sigma^2)$.

Estimator

Maximum Likelihood Estimator

Assumption: (MLE Estimator)

1. Random sampling
2. No Perfectly Collinearity
3. Zero Conditional Mean
4. Homoscedasticity
5. Gaussian Error

Theorem: (Maximum of Likelihood Estimator of OLS) Under the required assumption, the Maximum of Likelihood Estimator of the regression model is:

$$\hat{\beta} = (X'X)^{-1}X'y = (\sum_{i=1}^{n} x_i x_i')^{-1}\sum_{i=1}^{n} x_i y_i \tag{13}$$

$$\hat{\sigma}^2 = \hat{e}'\hat{e}/n = (y - X\hat{\beta})'(y - X\hat{\beta})/n = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2/n$$

Proof:

By definition we have $\hat{\beta}$ is maximizing $ln(L(\beta, \sigma^2|X)) = \sum_{i=1}^{n} log(f(X_i|\beta, \sigma^2))$. When we assume that the Gaussian error is true, we have $ln(L(\beta, \sigma^2|X)) = \sum_{i=1}^{n}(-\frac{1}{2}ln(2\pi) - \frac{1}{2}ln(\sigma^2) - \frac{(y_i - x_i'\beta)^2}{2\sigma^2})$. Now take the first order condition, we have $\sum_{i=1}^{n} 2x_i(y_i - x_i'\beta) = 0$, which will give us $\hat{\beta} = (\sum_{i=1}^{n} x_i x_i')^{-1} \sum_{i=1}^{n} x_i y_i$. Similarly take the first order condition of σ^2, we have $\sum_{i=1}^{n}(-\frac{1}{2\sigma^2} + \frac{(y_i - x_i'\beta)^2}{2(\sigma^2)^2}) = 0$, which will give us $\hat{\sigma}^2 = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2/n$. □

Least Square Estimator

Assumption: (OLS Estimator)

1. Random sampling
2. No Perfectly Collinearity

Theorem: (OLS Estimator) Under the required assumption, the OLS Estimator is:

$$\hat{\beta} = (X'X)^{-1}X'y = (\sum_{i=1}^{n} x_i x_i')^{-1} \sum_{i=1}^{n} x_i y_i \tag{14}$$

Proof:

By definition, the OLS estimator is minimizing $\frac{1}{n}(y - Xb)'(y - Xb)$. Taking the first order condition, we have $X'(y - Xb) = 0$. Suppose $X'X$ is reversible, then we have $\hat{\beta} = (X'X)^{-1}X'y$. □

Definition: (Prediction) Under the required assumption, the Prediction of the dependent variable is the estimator of $E[y|X]$, defined as:

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y \tag{15}$$

Definition: (Residual) Under the required assumption, the Residual of the estimation is defined as:

$$\hat{e} = y - \hat{y} = y - X\hat{\beta} = y - X(X'X)^{-1}X'y \tag{16}$$

Definition: (Projection Matrix) The Projection Matrix is Defined as:

$$P_X = X(X'X)^{-1}X' \tag{17}$$

Definition: (Orthogonal Projection Matrix) The Orthogonal Projection Matrix is Defined as:

$$M_X = I - X(X'X)^{-1}X' \tag{18}$$

Leverage

Definition: (Leverage) The Leverage of the estimation is defined as $h_{ii} = x_i'(X'X)^{-1}x_i$.
Definition: (Influence) The predict estimator is defined as $\hat{\beta}_{-i} = \hat{\beta} - (1 - h_{ii})^{-1}(X'X)^{-1}x_i\hat{e}_i$, and we define the prediction residual as $\tilde{e}_i = y_i - x_i'\hat{\beta}_{-i} = \hat{e}_i/(1 - h_{ii})$.

Note: $x_i'\hat{\beta} - x_i'\hat{\beta}_{-i} = h_{ii}\tilde{e}_i$.

General Properties of the Estimation

Theorem: (Properties of the Estimator and Residual) Under Assumption 1 and 2, the OLS Estimator and the Residual has the following properties:

1. $\hat{y} = P_X y$, and $\hat{e} = M_X e = M_X y$
2. $\hat{e}'\hat{e} = e'M_X e = y'M_X y$
3. $X'\hat{e} = 0$ and $\hat{y}'\hat{e} = 0$
4. If the independent variables include constant, i.e. $x_1 = \iota$, then $\sum_{i=1}^{n} \hat{e} = 0$, and $\bar{y} = \bar{\hat{y}}$

Proof:

1. First two can be shown by definition. We only want to show that $\hat{e}'\hat{e} = e'M_Xe = y'M_Xy$. This is because $M_Xy = M_X(X\beta + e))$ and $M_XX = 0$. Note that $\hat{e}'\hat{e} = (M_Xe)'M_Xe = e'M_Xe$.
2. The third equation is exactly the first order condition. $X'(y - X\hat{\beta}) = X'\hat{e} = 0$ and $\hat{y}'\hat{e} = (X\hat{\beta})'\hat{e} = 0$.
3. The forth argument comes from the first vector of equation 3. Since $\sum_{i=1}^n \hat{e} = 0$, we have $\sum_{i=1}^n \hat{e} = \sum_{i=1}^n (y_i - \hat{y}_i) = 0$, i.e. $\bar{y} = \bar{\hat{y}}$ □

Lemma:(Trace) For any two given matrix, $Trace(AB) = Trace(BA)$, as long as both traces exist.

Theorem: (Properties of the Projection Matrix) Under Assumption 1 and 2, the Projection Matrix has the following properties:

1. P_X is symmetric and idempotent, i.e. $P_X' = P_X$, and $P_XP_X = P_X$
2. If $X_1 = \iota$, then $P_X\iota = \iota$
3. $P_XX = X$
4. $P_\iota = \iota\iota'/n$
5. $P_\iota y = \bar{y}$
6. $Trace(P_X) = k$

Proof:

Most of the proof is trivial by definition. We only want to show equation 2 and 6. First we want to show equation 2., we have $P_X\iota = X(X'X)^{-1}X'\iota$, which is just regressing a constant on a set of random variables. Now prove equation 6. By the trace lemma, we have $Trace(P_X) = Trace(X(X'X)^{-1}X') = Trace((X'X)^{-1}(X'X)) = Trace(I_k) = k$. □

Theorem: (Properties of the Orthogonal Projection Matrix) Under Assumption 1 and 2, the Orthogonal Projection Matrix has the following properties:

1. M_X is symmetric and idempotent, i.e. $M_X' = M_X$, and $M_XM_X = M_X$
2. $M_XX = 0$
3. $M_\iota = I - \iota\iota'/n$
4. $Trace(M_X) = n - k$

Proof:

The first three proof is trivial. And we have $Trace(M_X) = Trace(I_n - P_X) = n - k$. □

Theorem: (Properties of the Leverage) Under Assumption 1 and 2, the Leverage has the following properties:

1. h_{ii} is the i-th element on the diagonal of P_X
2. $\sum_{i=1}^n h_{ii} = k$
3. $h_{ii} \in [0, 1]$

Proof:

By definition, h_{ii} is the i-th element on the diagonal of P_X. Since $Trace(P_X) = \sum_{i=1}^n h_{ii}$ we have $\sum_{i=1}^n h_{ii} = k$. We do not intend to show the last proof here. □

Special Cases

Theorem: (Special Regressor) The following statements are true:

1. When $k = 1$ and $X_1 = \iota$, $\hat{\beta} = \bar{y}$
2. When $k = 1$ and $X_1 = x$, $\hat{\beta} = \sum_{i=1}^n x_iy_i / \sum_{i=1}^n x_i^2$
3. When $k = 2$ and $X_1 = \iota$, $X_2 = x$, then $\hat{\beta}_1 = \bar{y} - \bar{x}\hat{\beta}_2$, and $\hat{\beta}_2 = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^n (x_i - \bar{x})^2$
4. (Transformations) When regress y on XC, the estimator is $\hat{\beta}^* = C^{-1}\hat{\beta}$, and $\hat{y}^* = \hat{y}$
5. (Transformations) When regress $a\iota + by$ on $X_1 = \iota$ and X_2, the estimator is $\hat{\beta}_1^* = a + b\hat{\beta}_1$, and $\hat{\beta}_2^* = b\hat{\beta}_2$

Proof:

1. The first two equations are trivial to prove.
2. Now prove the third equation. Since we have $\hat{y}'\hat{e}=0$, this implies $\bar{y}=\bar{\hat{y}}=\hat{\beta}_1+\bar{x}\hat{\beta}_2$. And $\hat{\beta}_2=\sum_{i=1}^n(x_i-\bar{x})(y_i-\bar{y})/\sum_{i=1}^n(x_i-\bar{x})^2$ comes from partitioned regression. This is shown in next part. Plug in the formula with dimension 1, we have $\beta_2=(X'M_\iota X)^{-1}X'M_\iota Y=[\sum_{i=1}^n(x_i-\bar{x})(x_i-\bar{x})']^{-1}\sum_{i=1}^n(x_i-\bar{x})(y_i-\bar{y})'$.
3. Now prove the transformations. Regressing y on XC, we have $\hat{\beta}^*=((XC)'(XC))^{-1}(XC)'y=(C'X'XC)^{-1}C'X'y$, then we have $\hat{\beta}^*=C^{-1}(X'X)^{-1}C'^{-1}C'X'y=C^{-1}\hat{\beta}$. And $\hat{y}^*=XC\hat{\beta}^*=XCC^{-1}\hat{\beta}=\hat{y}$.

   Now regress $a\iota+by$ on X_1 and X_2, we have $\hat{\beta}^*=(X'X)^{-1}X'(a\iota+by)=av+b\hat{\beta}$, where $v=(1,0,0,\dots,0)'$, which will give us what we need. □

Partitioned Regression

Partitioned Regression

Theorem: (Partitioned Regression) Suppose we see the regression model as $y=X_1\beta_1+X_2\beta_2+e$, then we have:

$$\hat{\beta}_1=(X_1'M_{X_2}X_1)^{-1}X_1'M_{X_2}y,\ \hat{\beta}_2=(X_2'M_{X_1}X_2)^{-1}X_2'M_{X_1}y \tag{19}$$

or

$$\hat{\beta}_1=((M_{X_2}X_1)'M_{X_2}X_1)^{-1}(M_{X_2}X_1)'y,\ \hat{\beta}_2=((M_{X_1}X_2)'M_{X_1}X_2)^{-1}(M_{X_1}X_2)'y \tag{20}$$

i.e. the regression of the residuals of y and X_1 on X_2.

Proof:

1. Remember we have the first order condition $X'(y-X_1\hat{\beta}_1-X_2\hat{\beta}_2)=0$. Note that $X=[X_1,X_2]$, so the first order condition can be partitioned into two equations. $X_1'(y-X_1\hat{\beta}_1-X_2\hat{\beta}_2)=0$ and $X_2'(y-X_1\hat{\beta}_1-X_2\hat{\beta}_2)=0$. This implies
$$\begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix}\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}=\begin{pmatrix} X_1'y \\ X_2'y \end{pmatrix} \tag{21}$$
   Now take the inverse of the left hand side we get the equation that we want. When $(X_1'M_{X_2}X_1)^{-1}=(X_1'X_1-X_1'X_2(X_2'X_2)^{-1}X_2'X_1)^{-1}$exists, we have
$$\begin{aligned}\hat{\beta}_1&=(X_1'M_{X_2}X_1)^{-1}\begin{pmatrix}1 & -X_1'X_2(X_2'X_2)^{-1}\end{pmatrix}\begin{pmatrix} X_1'y \\ X_2'y \end{pmatrix} \\ &=(X_1'M_{X_2}X_1)^{-1}X_1'(I-P_{X_2})y=(X_1'M_{X_2}X_1)^{-1}X_1'M_{X_2}y\end{aligned} \tag{22}$$
   When $(X_2'M_{X_1}X_2)^{-1}=(X_2'X_2-X_2'X_1(X_1'X_1)^{-1}X_1'X_2)^{-1}$ exists, we have the other half of the equation.
2. Now prove the same estimator is the result of doing the regression of the residuals of y and X_i on X_j. First regress $M_{X_1}y$ on X, we will get that by definition $M_{X_1}y=M_{X_1}X_1\beta_1+M_{X_1}X_2\beta_2+M_{X_1}e$. However, we know that $M_{X_1}X_1=0$. This implies that $M_{X_1}y=M_{X_1}X_2\beta_2+M_{X_1}e$ and hence the estimator $\hat{\beta}_2=((M_{X_1}X_2)'M_{X_1}X_2)^{-1}(M_{X_1}X_2)'M_{X_1}y$. □

Special Cases

Theorem: (Special Partitioned Regression) The following statements are true:

1. When $\hat{\beta}_1$ is a scaler and there is an intercept in X_2, then $\hat{\beta}_1=X_1'M_{X_2}y/(X_1'M_{X_2}X_1)$
2. Generally, when $X_1=\iota$, then the regression will pass the mean of the sample, i.e.

$$\hat{\beta}_1=\bar{y}-\bar{x}'\hat{\beta}_2=(\iota'M_{X_2}\iota)^{-1}\iota'M_{X_2}y \tag{23}$$

$$\hat{\beta}_2 = [\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})']^{-1} \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})' = (X_2' M_\iota X_2)^{-1} X_2' M_\iota y \tag{24}$$

Proof:

1. Plug in the formula from last theorem.
2. Since we have $\hat{y}'\hat{e} = 0$, this implies $\bar{y} = \bar{\hat{y}} = \hat{\beta}_1 + \bar{x}\hat{\beta}_2$. Plug in the formula from last theorem, we have $\hat{\beta}_2 = (X_2' M_\iota X_2)^{-1} X_2' M_\iota y = [\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})']^{-1} \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})'$. $\square$

R-Squared

Variation Partition

Definition:(Total Sum of Square) Total Sum of Square is defined as $SST = (y - \iota\bar{y})'(y - \iota\bar{y})$.

Definition:(Regression Sum of Square) Regression Sum of Square is defined as $SSR = (\hat{y} - \iota\bar{y})'(\hat{y} - \iota\bar{y}) = \hat{\beta}' X' M_\iota X\hat{\beta}$.

Definition:(Sum of Square Error) Sum of Square Error is defined as $SSE = \hat{e}'\hat{e} = \sum_{i=1}^{n} \hat{e}_i^2$.

Theorem: (Variation Partition) The following statements are true:

1. $y = P_X y + M_X y$
2. $SST = SSR + SSE$

Proof:

1. First equation is automatically true by definition.
2. $SST = (y - \iota\bar{y})'(y - \iota\bar{y})$, by $y = P_X y + M_X y$ we have $SST = (\hat{y} - \iota\bar{y} + \hat{e})'(\hat{y} - \iota\bar{y} + \hat{e}) = (\hat{y} - \iota\bar{y})'(\hat{y} - \iota\bar{y}) + \hat{e}'\hat{e}$ since we have $(\hat{y} - \iota\bar{y})'\hat{e} = \hat{e}'(\hat{y} - \iota\bar{y}) = 0$. This is because $(\hat{y} - \iota\bar{y})'\hat{e} = \hat{y}'\hat{e} - \iota\bar{y}'\hat{e} = 0 - 0 = 0$. $\square$

R-Squared

Definition:(R-Squared) R-Squared is defined as $R^2 = SSR/SST = 1 - SSE/SST$.

Theorem: (Properties of R-Squared) The following statements are true:

1. $R^2 = corr(y, \hat{y})^2$ for the sample
2. $R^2 \in [0, 1]$
3. When k increases, R-squared will always increase.

Proof:

it is trivial to show that $R^2 \in [0, 1]$. By definition we have $R^2 = SSR/SST = (\hat{y} - \iota\bar{y})'(\hat{y} - \iota\bar{y})/(y - \iota\bar{y})'(y - \iota\bar{y})$, where $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ and $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$. So we can rewrite $R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}$. Note that the numerator is

$$\begin{aligned}
\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 &= (\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(\hat{y}_i - \bar{y}))^2 = (\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(\hat{y}_i - \bar{y}) + \hat{e}_i(\hat{y}_i - \bar{y}))^2 \\
&= (\sum_{i=1}^{n}(\hat{y}_i - \bar{y} + \hat{e}_i)(\hat{y}_i - \bar{y}))^2 = (\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y}))^2
\end{aligned} \tag{25}$$

Hence $R^2 = corr(y, \hat{y})^2$ for the sample.

Now want to show that when k increases, R-squared will always increase. Consider an OLS regressing y on to $x_1, \ldots, x_k$, and suppose $\hat{\beta}_1, \ldots, \hat{\beta}_k$ minimize the SSE of the regression. Now suppose another x_{k+1} is added to the regression, If we plug in $\hat{\beta}_1, \ldots, \hat{\beta}_k, 0$ it will generate the R-squared before adding the variable. If we redo the OLS and get $\hat{\beta}_1^*, \ldots, \hat{\beta}_k^*, \hat{\beta}_{k+1}^*$, we will get the new R-squared. However, $\hat{\beta}_1^*, \ldots, \hat{\beta}_k^*, \hat{\beta}_{k+1}^*$ minimize the new SSE, and hence leading to a higher R-squared.

□

Adjusted R-Squared

Definition:(Adjusted R-Squared) Adjusted R-Squared is defined as:

$$R^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} \tag{26}$$

Properties of Estimator and Applications

General Small Sample Result

Assumption: (Small Sample Assumption)

1. Random sampling
2. No Perfectly Collinearity
3. Zero Conditional Mean, i.e. $E[e_i|x_i] = 0$

Theorem: (Small Sample Result) Under Assumption 1, 2 and 3, the following properties are true:

1. $\hat{\beta}$ is an unbiased estimator, i.e. $E[\hat{\beta}] = \beta$, and $E[\hat{e}] = 0$
2. $Var(\hat{\beta}|X) = (X'X)^{-1}X'\Sigma X(X'X)^{-1}$, where $\Sigma = E[ee'|X] = diag[\sigma^2(x_i)]$
3. $Var(\hat{e}|X) = M_X \Sigma M_X'$

   And when homoscedasticity is true, we have:
4. $Var(\hat{\beta}|X) = \sigma^2(X'X)^{-1}$
5. $Var(\hat{e}|X) = \sigma^2 M_X$
6. $E[\hat{e}_i^2|X] = \sigma^2(1-h_{ii})$

Proof:

1. $E[\hat{\beta}] = E[(X'X)^{-1}X'y] = E[(X'X)^{-1}X'X\beta] + E[(X'X)^{-1}X'e] = \beta + E[(X'X)^{-1}X'E[e|X] = \beta$.
2. $Var(\hat{\beta}|X) = Var((X'X)^{-1}X'y|X) = Var((X'X)^{-1}X'(X\beta+e)|X) = Var((X'X)^{-1}X'e|X) = (X'X)^{-1}X'Var(e|X)X(X'X)^{-1}$, where $Var(e|X) = \Sigma = E[ee'|X] = diag[\sigma^2(x_i)]$.
3. $Var(\hat{e}|X) = Var(M_X y|X) = Var(M_X e|X) = M_X \Sigma M_X'$.

   When homoscedasticity is true, we have:
4. $Var(\hat{\beta}|X) = (X'X)^{-1}X'\sigma^2 X(X'X)^{-1} = \sigma^2(X'X)^{-1}$
5. $Var(\hat{e}|X) = M_X \sigma^2 M_X' = \sigma^2 M_X$
6. Since by equation 5 we have $Var(\hat{e}|X) = M_X\sigma^2 M_X' = \sigma^2 M_X$. Now by definition h_{ii} is the i-th element on the diagonal of P_X, so $1 - h_{ii}$ is the i-th element on the diagonal of M_X, so we can write the i-th row of equation 5, which is $E[\hat{e}_i^2|X] = \sigma^2(1-h_{ii})$. □

Variance Estimation

Definition: (Heteroskedasticity variance estimator) When Assumption 1-3 are true and Heteroskedasticity is true, define the estimator of the variance of $\hat{\beta}$ as:

$$\hat{V}(\hat{\beta}|X) = (X'X)^{-1}X'SX(X'X)^{-1} \tag{27}$$

where $S = \hat{\Sigma} = diag[\hat{e}_i^2]$

Definition: (Homoscedasticity variance estimator) When Assumption 1-3 are true and Homoscedasticity is true, define the estimator of the variance of $\hat{e}$ as:

$$s^2 = \frac{\hat{e}'\hat{e}}{n-k} \tag{28}$$

Definition: (Standardized Residual) When Homoscedasticity is true, define the Standardized Residual as:

$$\bar{e}_i = \frac{\hat{e}}{\sqrt{1-h_{ii}}} \tag{29}$$

Definition: (Homoscedasticity variance estimator) When Homoscedasticity is true, define the estimator of the variance of e as:

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^n \bar{e}_i^2}{n} \tag{30}$$

Note: Under Homoscedasticity, $\hat{\sigma}^2$, s^2, and $\tilde{\sigma}^2$ are all estimators of σ^2, where the first is the MLE estimator, and the second and the third are generated because they are unbiased.

Definition: (Homoscedasticity variance estimator) When Homoscedasticity is true, define the estimator of the variance of $\hat{\beta}$ as:

$$\hat{V}(\hat{\beta}|X) = s^2(X'X)^{-1} \tag{31}$$

Theorem: (Expectation of the variance estimator) Under Assumption 1, 2, and 3, the following properties are true:

1. $E[\hat{V}(\hat{\beta}|X)] = Var(\hat{\beta}|X)$

   And when homoscedasticity is true, we have:

2. $E[s^2] = E[\tilde{\sigma}^2] = \sigma^2$, but $E[\hat{\sigma}^2] = (n-k)\sigma^2$

Proof:

1. $E[\hat{V}(\hat{\beta}|X)] = E[(X'X)^{-1}X'SX(X'X)^{-1}] = E[(X'X)^{-1}X'E[S|X]X(X'X)^{-1}]$ and
2. $E[\hat{e}'\hat{e}] = E[e'M_Xe|X] = E[Trace(e'M_Xe)|X] = E[Trac(M_Xe'e)|X] = Trace(M_XE[e'e|X]) = \sigma^2 Trace(M_X) = \sigma^2(n-k)$, so we have when homoscedasticity is true, we have $E[s^2] = E[\tilde{\sigma}^2] = E[\frac{\hat{e}'\hat{e}}{n-k}]$. $\square$

Gauss Markov Theorem

Efficient Estimator

Assumption: (Gauss Markov Assumption)

1. Random sampling
2. No Perfectly Collinearity
3. Zero Conditional Mean, i.e. $E[e_i|x_i] = 0$
4. Homoscedasticity
5. Gaussian Error

Theorem: (Gauss Markov Theorem) Under Assumption 1-5, OLS estimator is Best Linear Unbiased Estimator(BLUE).

Proof:

We want to show that there is no linear unbiased estimator that have a lower conditional variance. The conditional variance of any given estimator is $Var(\tilde{\beta}|X) = E[(\tilde{\beta}-\beta)'(\tilde{\beta}-\beta)|X]$, where $\tilde{\beta} = C'y$ is a linear estimator. It is also unbiased so $E[\tilde{\beta}] = E[C'(X\beta+e)] = C'X\beta$ implies that $C'X = I$. So $E[(\tilde{\beta}-\beta)'(\tilde{\beta}-\beta)|X] = C'E[ee'|X]C = \sigma^2 C'C$.

Now we have $C'C = (C - X(X'X)^{-1} + X(X'X)^{-1})'(C - X(X'X)^{-1} + X(X'X)^{-1})$, which can be written as $(C - X(X'X)^{-1})'(C - X(X'X)^{-1}) + (X'X)^{-1}$. Because $(C - X(X'X)^{-1})'X(X'X)^{-1} = (CX - I)(X'X)^{-1} = 0$. Then since the first part of $C'C$ is a positive semi-definite matrix, we have $C'C \geq (X'X)^{-1}$, which shows that there is no linear unbiased estimator that have a lower conditional variance. $\square$

Claim: (WLS Theorem) Under Assumption 1, 2, 3, and Heteroskedasticity, OLS estimator is not the Best Linear Unbiased Estimator(BLUE), instead, The BLUE is:

$$\hat{\beta}_W = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y \tag{32}$$

Note: Under homoscedasticity WLS will give the same estimator as OLS.

Small Sample Distribution Result

Assumption: (Small Sample Distribution Assumption)

1. Random sampling
2. No Perfectly Collinearity
3. Zero Conditional Mean, i.e. $E[e_i|x_i] = 0$
4. Homoscedasticity
5. Gaussian Error

Theorem: (Conditional Distribution) Under Assumption 1-5, the following statement are true:

1. $\hat{\beta}|X \sim N(\beta, \sigma^2(X'X)^{-1})$
2. $\hat{e}|X \sim N(0, \sigma^2 M_X)$
3. $\hat{\beta}$ is independent to $\hat{e}$
4. $(n-k)s^2/\sigma^2 \sim \chi^2(n-k)$
5. $\hat{\beta}$ is independent to s^2
6. $T_j|X = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2[(X'X)^{-1}]_{jj}}}|X \sim N(0,1)$
7. $\hat{T}_j|X = \frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2[(X'X)^{-1}]_{jj}}}|X \sim T(n-k)$
8. When C is a $1 \times k$ vector, we have $\hat{T}'|X = \frac{C\hat{\beta} - C\beta}{\sqrt{s^2 C(X'X)^{-1}C'}}|X \sim T(n-k)$
9. When R is a $J \times k$ matrix, we have $F|X = \frac{(R(\hat{\beta}-\beta))'(R(X'X)^{-1}R')^{-1}(R(\hat{\beta}-\beta))/J}{\sigma^2}|X \sim \chi^2(J)/J$
10. When R is a $J \times k$ matrix, we have $\hat{F}|X = \frac{(R(\hat{\beta}-\beta))'(R(X'X)^{-1}R')^{-1}(R(\hat{\beta}-\beta))/J}{s^2}|X \sim F(J, n-k)$

Proof:

1. $\hat{\beta}|X = (X'X)^{-1}X'y|X = (X'X)^{-1}X'(X\beta+e)|X \sim N(\beta, \sigma^2(X'X)^{-1})$, by assumption $e|X \sim N(0, \sigma^2)$.
2. $\hat{e}|X = M_X e|X \sim N(0, \sigma^2 M_X)$.
3. Now want to show that $Cov(\hat{\beta}, \hat{e}) = 0$. $Cov(\hat{\beta}, \hat{e}) = E[(\hat{\beta}-\beta)\hat{e}'|X] = E[(X'X)^{-1}X'e(M_X e)'|X] = E[(X'X)^{-1}X'ee'M_X|X]$. But we have $E[ee'|X] = \sigma^2$, so $Cov(\hat{\beta}, \hat{e}) = \sigma^2(X'X)^{-1}X'M_X = 0$, since $M_X X = 0$. Under normality, $\hat{\beta}$ is independent to $\hat{e}$.
4. $(n-k)s^2/\sigma^2 = \frac{1}{\sigma^2}e'M_X'M_X e = (\frac{e}{\sigma})'M_X'M_X(\frac{e}{\sigma}) = (\frac{e}{\sigma})'M_X(\frac{e}{\sigma})$. We know that $\frac{e}{\sigma}|X \sim N(0, I_n)$. Now we take the spectral decomposition of M_X. We have $M_X = H\Lambda H'$, where

$$\Lambda = \begin{pmatrix} I_{n-k} & 0 \\ 0 & 0 \end{pmatrix} \tag{33}$$

Note that the eigenvalues of M_X are either 0 or 1. So $\sum_{i=1}^{n}\lambda_i = Trace(M_X) = n-k$. We also have $H'H = HH' = I_n$ and $H^{-1} = H'$ because M_X is a symmetric and idempotent matrix. Then we define $(\frac{e}{\sigma})'M_X(\frac{e}{\sigma}) = z'\Lambda z$, and we have $z = H'(\frac{e}{\sigma})|X \sim N(0, H'I_nH) = N(0, I_n)$

$$z'\Lambda z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}' \begin{pmatrix} I_{n-k} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = z_1 I_{n-k} z_1 = \sum_{i=1}^{n-k} z_{1i}^2 \sim \chi^2(n-k) \tag{34}$$

5. Since $\hat{\beta}$ is independent to $\hat{e}$, we have $\hat{\beta}$ is independent to s^2, which is a function of $\hat{e}$.
6. From above this is true by definition.
7. From above this is true by definition.
8. By linear combination of normal distribution, we have $C(\hat{\beta}-\beta)|X \sim N(0, \sigma^2 C(X'X)^{-1}C')$. So this is true by the definition of T distribution.
9. From above this is true by definition.
10. From above this is true by definition. □

Theorem: (Partitioned Regression) Suppose we see the regression model as $Y = X_1\beta_1 + X_2\beta_2 + e$. Under Assumption 1-5, we have:

1. $\hat{\beta}_1|X \sim N(\beta_1, \sigma^2(X_1'X_1 - X_1'X_2(X_2'X_2)^{-1}X_2'X_1)^{-1})$
2. $\hat{\beta}_2|X \sim N(\beta_2, \sigma^2(X_2'X_2 - X_2'X_1(X_1'X_1)^{-1}X_1'X_2)^{-1})$

Proof:

By argument 1 from the last theorem, we have $\hat{\beta}|X \sim N(\beta, \sigma^2(X'X)^{-1})$. If we write $X = (X_1, X_2)$, we can use the partition of matrix and we will get:

$$X'X = \begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix} \tag{35}$$

When $(X_1'M_{X_2}X_1)^{-1} = (X_1'X_1 - X_1'X_2(X_2'X_2)^{-1}X_2'X_1)^{-1}$exists, we have what we want to show. Suppose $(X_2'X_2 - X_2'X_1(X_1'X_1)^{-1}X_1'X_2)^{-1}$ exists, we can prove the other half. □

Large Sample Theory

Theory Under Heteroscedasticity

Assumption: (Large Sample Distribution Assumption with Heteroscedasticity)

1. Random sampling
2. No Perfectly Collinearity
3. Zero Correlation, i.e. $E[x_i e_i] = 0$
4. $E[x_i x_i' e_i^2] = \Omega < +\infty$
5. $E[x_i x_i'] = Q_{xx} < +\infty$ and it is positive definite

Theorem: (Consistency) Under Assumption 1-5, suppose we have large sample, then the OLS estimator is consistent.

Proof:

We want to show that $\hat{\beta} \to^p \beta$. We have $\hat{\beta} = (X'X)^{-1}X'y = \beta + (X'X/n)^{-1}(X'e/n)$, where $(X'X/n)^{-1} = (\sum_{i=1}^{n} x_i x_i'/n)^{-1} \to^p Q_{xx}^{-1}$, by the law of large number, and $(X'e/n) = (\sum_{i=1}^{n} x_i e_i/n) \to^p E[x_i e_i] = 0$ also by the law of large number. □

Theorem: (Asymptotic Result) Under Assumption 1-5, suppose we have large sample, then the following results are true:

1. $\sqrt{n}(\hat{\beta}-\beta)|X \to^d N(0, Q_{xx}^{-1}\Omega Q_{xx}^{-1})$

2. $lim_p nV(\hat{\beta}|X) = Q_{xx}^{-1}\Omega Q_{xx}^{-1}$
3. $lim_p n\hat{V}(\hat{\beta}|X) = Q_{xx}^{-1}\Omega Q_{xx}^{-1}$
4. $\hat{T}_j|X = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{V}(\hat{\beta}|X)_{jj}}}|X \rightarrow^d N(0,1)$
5. When C is a $1 \times k$ vector, we have $\hat{T}'|X = \frac{C\hat{\beta} - C\beta}{\sqrt{C\hat{V}(\hat{\beta}|X)C'}}|X \rightarrow^d N(0,1)$
6. When R is a $J \times k$ matrix, we have $F|X = (R(\hat{\beta}-\beta))'(RV(\hat{\beta}|X)R')^{-1}(R(\hat{\beta}-\beta))/J|X \rightarrow^d \chi^2(J)/J$
7. When R is a $J \times k$ matrix, we have $\hat{F}|X = (R(\hat{\beta}-\beta))'(R\hat{V}(\hat{\beta}|X)R')^{-1}(R(\hat{\beta}-\beta))/J|X \rightarrow^d \chi^2(J)/J$
8. Generally, suppose $g(.)$ is a function system with J equations, $\sqrt{n}(g(\hat{\beta}) - g(\beta)) \rightarrow^d N(0, G'Q_{xx}^{-1}\Omega Q_{xx}^{-1}G)$, where $G = \partial g(\beta)/\partial\beta|_{\hat{\beta}}$
9. Generally, suppose $g(.)$ is a function system with J equations, $\hat{W} = (g(\hat{\beta}) - g(\beta))'(G'\hat{V}(\hat{\beta}|X)G)^{-1}(g(\hat{\beta}) - g(\beta))/J \rightarrow^d \chi^2(J)/J$, where $G = \partial g(\beta)/\partial\beta|_{\hat{\beta}}$

Proof:

1. $\sqrt{n}(\hat{\beta}-\beta) = \sqrt{n}((X'X/n)^{-1}(X'e/n))$ where $(X'X/n)^{-1} \rightarrow^p Q_{XX}^{-1}$ by the law of large number and $\sqrt{n}(X'e/n) \rightarrow^d N(0,\Omega)$ by the central limit theorem. Combine them we get $\sqrt{n}(\hat{\beta}-\beta)|X \rightarrow^d N(0, Q_{xx}^{-1}\Omega Q_{xx}^{-1})$.
2. Note that $V(\hat{\beta}|X) = V((X'X)^{-1}X'e|X)$, so $nV(\hat{\beta}|X) = (X'X/n)^{-1}E[X'ee'X/n|X](X'X/n)^{-1}$ Then $(X'X/n)^{-1} \rightarrow^p Q_{XX}^{-1}$, and $E[X'ee'X/n|X] \rightarrow^p \Omega$. Combine them we have $lim_p nV(\hat{\beta}|X) = Q_{xx}^{-1}\Omega Q_{xx}^{-1}$.
3. Note that $n\hat{V}(\hat{\beta}|X) = (X'X/n)^{-1}(X'SX/n)(X'X/n)^{-1}$. Then $(X'X/n)^{-1} \rightarrow^p Q_{XX}^{-1}$, and $X'SX/n = \frac{1}{n}\sum_{i=1}^n x_i x_i' \hat{e}_i^2 \rightarrow^p \Omega$ by the law of large number. Combine them we have $lim_p nV(\hat{\beta}|X) = Q_{xx}^{-1}\Omega Q_{xx}^{-1}$.
4. $\hat{T}_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{V}(\hat{\beta}|X)_{jj}}}$, and since equation 1 and 3 are true, we can combine them and conclude that $\hat{T}_j|X = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{V}(\hat{\beta}|X)_{jj}}}|X \rightarrow^d N(0,1)$.
5. $\sqrt{n}(C\hat{\beta} - C\beta) = \sqrt{n}C((X'X/n)^{-1}(X'e/n)) \rightarrow^d N(0, CQ_{xx}^{-1}\Omega Q_{xx}^{-1}C')$, and $nC\hat{V}(\hat{\beta}|X)C' \rightarrow^p CQ_{xx}^{-1}\Omega Q_{xx}^{-1}C'$. Combine them we will get $\hat{T}'|X = \frac{C\hat{\beta} - C\beta}{\sqrt{C\hat{V}(\hat{\beta}|X)C'}}|X \rightarrow^d N(0,1)$.
6. We have $F = (\sqrt{n}R(\hat{\beta}-\beta))'(nRV(\hat{\beta}|X)R')^{-1}(\sqrt{n}R(\hat{\beta}-\beta))/J$. Now $\sqrt{n}R(\hat{\beta}-\beta) \rightarrow^d N(0, RQ_{xx}^{-1}\Omega Q_{xx}^{-1}R')$, and $nRV(\hat{\beta}|X)R' \rightarrow^p RQ_{xx}^{-1}\Omega Q_{xx}^{-1}R'$. Combine them we have $F|X = (R(\hat{\beta}-\beta))'(RV(\hat{\beta}|X)R')^{-1}(R(\hat{\beta}-\beta))/J|X \rightarrow^d \chi^2(J)/J$.
7. We have $\hat{F} = (\sqrt{n}R(\hat{\beta}-\beta))'(nR\hat{V}(\hat{\beta}|X)R')^{-1}(\sqrt{n}R(\hat{\beta}-\beta))/J$. Now $\sqrt{n}R(\hat{\beta}-\beta) \rightarrow^d N(0, RQ_{xx}^{-1}\Omega Q_{xx}^{-1}R')$, and $nR\hat{V}(\hat{\beta}|X)R' \rightarrow^p RQ_{xx}^{-1}\Omega Q_{xx}^{-1}R'$. Combine them we have $\hat{F}|X = (R(\hat{\beta}-\beta))'(R\hat{V}(\hat{\beta}|X)R')^{-1}(R(\hat{\beta}-\beta))/J|X \rightarrow^d \chi^2(J)/J$.
8. By equation 1 we have already shown that $\sqrt{n}(\hat{\beta}-\beta)|X \rightarrow^d N(0, Q_{xx}^{-1}\Omega Q_{xx}^{-1})$. Use delta method and we get what we want to show.
9. We only need to show that $nG'\hat{V}(\hat{\beta}|X)G \rightarrow^p G'Q_{xx}^{-1}\Omega Q_{xx}^{-1}G$, which is true from what we have already shown before. □

Theory Under Homoscedasticity

Assumption: (Large Sample Distribution Assumption with Homoscedasticity)

1. Random sampling
2. No Perfectly Collinearity
3. Zero Correlation, i.e. $E[x_i e_i] = 0$
4. $E[x_i x_i' e_i^2] = \Omega < +\infty$
5. $E[x_i x_i'] = Q_{xx} < +\infty$ and it is positive definite
6. Homoscedasticity

Theorem: (Asymptotic Result with Homoscedasticity) Under Assumption 1-6, suppose we have large sample and Homoscedasticity is true, we have:

1. $\sqrt{n}(\hat{\beta}-\beta)|X \rightarrow^d N(0, \sigma^2 Q_{xx}^{-1})$
2. $lim_p n\sigma^2(X'X)^{-1} = \sigma^2 Q_{xx}^{-1}$
3. $lim_p ns^2(X'X)^{-1} = \sigma^2 Q_{xx}^{-1}$

4. $\hat{T}_j|X = \frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2[(X'X)^{-1}]_{jj}}}|X \rightarrow^d N(0,1)$
5. When C is a $1 \times k$ vector, we have $\hat{T}'|X = \frac{C\hat{\beta} - C\beta}{\sqrt{s^2[C(X'X)^{-1}C']}}|X \rightarrow^d N(0,1)$
6. When R is a $J \times k$ matrix, we have $F|X = \frac{(R(\hat{\beta}-\beta))'(R(X'X)^{-1}R')^{-1}(R(\hat{\beta}-\beta))/J}{\sigma^2}|X \rightarrow^d \chi^2(J)/J$
7. When R is a $J \times k$ matrix, we have $\hat{F}|X = \frac{(R(\hat{\beta}-\beta))'(R(X'X)^{-1}R')^{-1}(R(\hat{\beta}-\beta))/J}{s^2}|X \rightarrow^d \chi^2(J)/J$
8. Generally, suppose $g(.)$ is a function system with J equations, $\sqrt{n}(g(\hat{\beta}) - g(\beta)) \rightarrow^d N(0, \sigma^2 G' Q_{xx}^{-1} G)$, where $G = \partial g(\beta)/\partial \beta|_{\hat{\beta}}$
9. Generally, suppose $g(.)$ is a function system with J equations, $\hat{W} = \frac{(g(\hat{\beta})-g(\beta))'(G'(X'X)^{-1}G)^{-1}(g(\hat{\beta})-g(\beta))/J}{s^2} \rightarrow^d \chi^2(J)/J$, where $G = \partial g(\beta)/\partial \beta|_{\hat{\beta}}$

Proof:

1. $\sqrt{n}(\hat{\beta} - \beta) = \sqrt{n}((X'X/n)^{-1}(X'e/n))$ where $(X'X/n)^{-1} \rightarrow^p Q_{xx}^{-1}$ by the law of large number and $\sqrt{n}(X'e/n) \rightarrow^d N(0, \Omega) = N(0, \sigma^2 Q_{xx})$ by the central limit theorem. Combine them we get $\sqrt{n}(\hat{\beta} - \beta)|X \rightarrow^d N(0, \sigma^2 Q_{xx}^{-1})$.
2. Note that $V(\hat{\beta}|X) = V((X'X)^{-1}X'e|X) = \sigma^2 (X'X)^{-1}$, so $nV(\hat{\beta}|X) = (X'X/n)^{-1}X'E[ee'/n|X]X(X'X/n)^{-1}$ Then $(X'X/n)^{-1} \rightarrow^p Q_{XX}^{-1}$. Combine them we have $lim_p nV(\hat{\beta}|X) = \sigma^2 Q_{xx}^{-1} Q_{xx} Q_{xx}^{-1} = \sigma^2 Q_{xx}^{-1}$.
3. Note that $n\hat{V}(\hat{\beta}|X) = (X'X/n)^{-1}s^2$. Then $(X'X/n)^{-1} \rightarrow^p Q_{XX}^{-1}$, and $s^2 = \frac{\hat{e}'\hat{e}}{n-k} \rightarrow^p \sigma^2$ by the law of large number. Combine them we have $lim_p ns^2(X'X)^{-1} = \sigma^2 Q_{xx}^{-1}$.
4. $\hat{T}_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2[(X'X)^{-1}]_{jj}}}$, and since equation 1 and 3 are true, we can combine them and conclude that $\hat{T}_j|X = \frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2[(X'X)^{-1}]_{jj}}}|X \rightarrow^d N(0,1)$.
5. $\sqrt{n}(C\hat{\beta} - C\beta) = \sqrt{n}C((X'X/n)^{-1}(X'e/n)) \rightarrow^d N(0, C\sigma^2 Q_{xx}^{-1}C')$, and $ns^2[C(X'X)^{-1}C'] \rightarrow^p C\sigma^2 Q_{xx}^{-1}C'$. Combine them we will get $\hat{T}'|X = \frac{C\hat{\beta} - C\beta}{\sqrt{s^2[C(X'X)^{-1}C']}}|X \rightarrow^d N(0,1)$.
6. We have $F = \frac{(R(\hat{\beta}-\beta))'(R(X'X)^{-1}R')^{-1}(R(\hat{\beta}-\beta))/J}{\sigma^2}$. Now $\sqrt{n}R(\hat{\beta} - \beta) \rightarrow^d N(0, R\sigma^2 Q_{xx}^{-1}R')$. So we have $F|X = \frac{(R(\hat{\beta}-\beta))'(R(X'X)^{-1}R')^{-1}(R(\hat{\beta}-\beta))/J}{\sigma^2}|X \rightarrow^d \chi^2(J)/J$.
7. We have $\hat{F} = \frac{(R(\hat{\beta}-\beta))'(R(X'X)^{-1}R')^{-1}(R(\hat{\beta}-\beta))/J}{s^2}$. Now $\sqrt{n}R(\hat{\beta} - \beta) \rightarrow^d N(0, R\sigma^2 Q_{xx}^{-1}R')$, and $nRs^2 Q_{xx}^{-1}R' \rightarrow^p R\sigma^2 Q_{xx}^{-1}R'$. Combine them we have $\hat{F}|X = \frac{(R(\hat{\beta}-\beta))'(R(X'X)^{-1}R')^{-1}(R(\hat{\beta}-\beta))/J}{s^2}|X \rightarrow^d \chi^2(J)/J$.
8. By equation 1 we have already shown that $\sqrt{n}(\hat{\beta} - \beta)|X \rightarrow^d N(0, \sigma^2 Q_{xx}^{-1})$. Use delta method and we get what we want to show.
9. We only need to show that $nG's^2 Q_{xx}^{-1}G \rightarrow^p G'\sigma^2 Q_{xx}^{-1}G$, which is true from what we have already shown before. □

Theorem: (Partitioned Regression) Suppose we see the regression model as $Y = X_1\beta_1 + X_2\beta_2 + e$. Under Assumption 1-6, suppose we have large sample and Homoscedasticity is true, we have:

1. $\sqrt{n}(\hat{\beta}_1 - \beta_1)|X \rightarrow^d N(0, \sigma^2(Q_{11} - Q_{12}Q_{22}^{-1}Q_{21})^{-1})$
2. $\sqrt{n}(\hat{\beta}_2 - \beta_2)|X \rightarrow^d N(0, \sigma^2(Q_{22} - Q_{21}Q_{11}^{-1}Q_{12})^{-1})$

Proof:

By argument 1 from the last theorem, we have $\sqrt{n}(\hat{\beta} - \beta)|X \rightarrow^d N(0, \sigma^2 Q_{xx}^{-1})$. If we write $X = (X_1, X_2)$, we can use the partition of matrix and we will get:

$$Q_{xx} = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \tag{36}$$

When $(X_{11} - Q_{12}(Q_{22})^{-1}Q_{21})^{-1}$exists, we have what we want to show. Suppose $(Q_{22} - Q_{21}Q_{11}^{-1}Q_{12})^{-1}$ exists, we can prove the other half. □

Hypothesis Test

Assumption

Assumption: (Small Sample)

1. Random sampling
2. No Perfectly Collinearity
3. Zero Conditional Mean, i.e. $E[e_i|x_i] = 0$
4. Homoscedasticity
5. Gaussian Error

Assumption: (Large Sample)

1. Random sampling
2. No Perfectly Collinearity
3. Zero Correlation, i.e. $E[x_i e_i] = 0$
4. $E[x_i x_i' e_i^2] = \Omega < +\infty$
5. $E[x_i x_i'] = Q_{xx} < +\infty$ and it is positive definite

Assumption: (Large Sample with Homoscedasticity)

1. Random sampling
2. No Perfectly Collinearity
3. Zero Correlation, i.e. $E[x_i e_i] = 0$
4. $E[x_i x_i' e_i^2] = \Omega < +\infty$
5. $E[x_i x_i'] = Q_{xx} < +\infty$ and it is positive definite
6. Homoscedasticity

T test

Method: (Test with Small Sample) Under the Assumption about small sample, we use the T estimator to do Hypothesis Test for $H_0 : \beta = \beta_0$, and $H_1 : \beta \neq \beta_0$, i.e. reject if $\hat{T} \notin [-T_{\alpha/2}, T_{\alpha/2}]$, where $\hat{T}$ is defined as:

$$\hat{T}_j|X = \frac{\hat{\beta}_j - \beta_{0j}}{\sqrt{s^2[(X'X)^{-1}]_{jj}}}|X \sim T(n-k) \tag{37}$$

Method: (Test with Large Sample) Under the Assumption about large sample and heteroskedasticity, we use the T estimator to do Hypothesis Test for $H_0 : \beta = \beta_0$, and $H_1 : \beta \neq \beta_0$, i.e. reject if $\hat{T} \notin [-T_{\alpha/2}, T_{\alpha/2}]$, where $\hat{T}$ is defined as:

$$\hat{T}_j|X = \frac{\hat{\beta}_j - \beta_{0j}}{\sqrt{\hat{V}(\hat{\beta}|X)_{jj}}}|X \to^d N(0,1) \tag{38}$$

Method: (Test with Large Sample and Homoscedasticity) Under the Assumption about large sample and homoscedasticity, we use the T estimator to do Hypothesis Test for $H_0 : \beta = \beta_0$, and $H_1 : \beta \neq \beta_0$, i.e. reject if $\hat{T} \notin [-T_{\alpha/2}, T_{\alpha/2}]$, where $\hat{T}$ is defined as:

$$\hat{T}_j|X = \frac{\hat{\beta}_j - \beta_{0j}}{\sqrt{s^2[(X'X)^{-1}]_{jj}}}|X \to^d N(0,1) \tag{39}$$

Theorem: (Unbiased and Consistent T-Test) The T-Test described above is unbiased under small sample assumption, and consistent under large sample assumption.

Proof:

1. Under small sample assumptions, we want to show that T-test is unbiased. Suppose the true value is β, instead of β_0. Then the T statistic is $T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2[(X'X)^{-1}]_{jj}}} + \frac{\beta_j - \beta_{0j}}{\sqrt{s^2[(X'X)^{-1}]_{jj}}}$, where the first part of the equation is defined as $T_0 = \frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2[(X'X)^{-1}]_{jj}}} \sim T(n-k)$. Under $H_0 : \beta_j = \beta_{0j}$, the second term is negative so we have $T = T_0$, and $P(|T| > t_{\alpha/2}) < \alpha$. Under $H_1 : \beta_j \neq \beta_{0j}$, we have $T \neq T_0$, and $P(|T| > t_{\alpha/2}) > \alpha$. So This test is unbiased.
2. Under large sample assumptions, and under $H_1 : \beta_j \neq \beta_{0j}$, we have :

$$|T| = |\frac{\hat{\beta}_j - \beta_{0j}}{\sqrt{\hat{V}(\hat{\beta}|X)_{jj}}}| = |\frac{\beta_j + (X'X)^{-1}(X'e) - \beta_{0j}}{\sqrt{\hat{V}(\hat{\beta}|X)_{jj}}}| = |\frac{\beta_j - \beta_{0j}}{\sqrt{\hat{V}(\hat{\beta}|X)_{jj}}} + \frac{(X'X)^{-1}(X'e)}{\sqrt{\hat{V}(\hat{\beta}|X)_{jj}}}| \tag{40}$$

where the first term goes to infinity when $H_1 : \beta_j \neq \beta_{0j}$ is true. Since the second term goes to a standard normal distribution, we have $P(|T| > z_{\alpha/2}) \rightarrow^p 1$, i.e. the test is constant. □

T Test with General Linear Restriction

Method: (Test with Small Sample) Under the Assumption about small sample, we use the linear combined T estimator to do Hypothesis Test for $H_0 : C\beta - r = 0$, and $H_1 : C\beta - r \neq 0$, where C is a $1 \times k$ vector, i.e. reject if $\hat{T}' \notin [-T_{\alpha/2}, T_{\alpha/2}]$, where $\hat{T}'$ is defined as:

$$\hat{T}'|X = \frac{C\hat{\beta} - r}{\sqrt{s^2 C(X'X)^{-1}C'}}|X \sim T(n-k) \tag{41}$$

Method: (Test with Large Sample) Under the Assumption about large sample and heteroscedasticity, we use the linear combined T estimator to do Hypothesis Test for $H_0 : C\beta - r = 0$, and $H_1 : C\beta - r \neq 0$, where C is a $1 \times k$ vector, i.e. reject if $\hat{T}' \notin [-N_{\alpha/2}, N_{\alpha/2}]$, where $\hat{T}'$ is defined as:

$$\hat{T}'|X = \frac{C\hat{\beta} - r}{\sqrt{C\hat{V}(\hat{\beta}|X)C'}}|X \rightarrow^d N(0,1) \tag{42}$$

Method: (Test with Large Sample and Homoscedasticity) Under the Assumption about large sample and homoscedasticity, we use the linear combined T estimator to do Hypothesis Test for $H_0 : C\beta - r = 0$, and $H_1 : C\beta - r \neq 0$, where C is a $1 \times k$ vector, i.e. reject if $\hat{T}' \notin [-N_{\alpha/2}, N_{\alpha/2}]$, where $\hat{T}'$ is defined as:

$$\hat{T}'|X = \frac{C\hat{\beta} - r}{\sqrt{s^2[C(X'X)^{-1}C']}}|X \rightarrow^d N(0,1) \tag{43}$$

F test

Method: (Test with Small Sample) Under the Assumption about small sample, we use the F estimator to do Hypothesis Test for $H_0 : R\beta - r = 0$, and $H_1 : R\beta - r \neq 0$, where R is a $J \times k$ vector, i.e. reject if $\hat{F} \in [F_\alpha, +\infty]$, where $\hat{F}$ is defined as:

$$\hat{F}|X = \frac{(R\hat{\beta} - r)'(R(X'X)^{-1}R')^{-1}(R\hat{\beta} - r))/J}{s^2}|X \sim F(J, n-k) \tag{44}$$

Theorem: (Alternative Derivation of F Statistic) Under the small sample assumptions, suppose we have $SSE_U = (Y - X\hat{\beta})'(Y - X\hat{\beta})$, and $SSE_R = (Y - X\tilde{\beta})'(Y - X\tilde{\beta})$ where $\tilde{\beta} = \hat{\beta} - (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}(R\hat{\beta} - r)$, then we have:

$$\hat{F} = \frac{(SSE_R - SSE_U)/J}{SSE_U/(n-k)} \tag{45}$$

Proof:

Under the small sample assumptions, we have:

$$\hat{F} = \frac{(SSE_R - SSE_U)/J}{SSE_U/(n-k)} = \frac{((y - X\tilde{\beta})'(y - X\tilde{\beta}) - (y - X\hat{\beta})'(y - X\hat{\beta}))/J}{SSE_U/(n-k)} \tag{46}$$
$$= \frac{1}{Js^2}((y - X\tilde{\beta})'(y - X\tilde{\beta}) - (y - X\hat{\beta})'(y - X\hat{\beta}))$$
$$= \frac{1}{Js^2}(y'y - \tilde{\beta}'X'y - y'X\tilde{\beta} + \tilde{\beta}'X'X\tilde{\beta} - y'y + \hat{\beta}'X'y + y'X\hat{\beta} - \hat{\beta}'X'X\hat{\beta})$$
$$= \frac{1}{Js^2}(-(\tilde{\beta} - \beta)'X'y - y'X(\tilde{\beta} - \beta) + \tilde{\beta}'X'X\tilde{\beta} + (\hat{\beta} - \beta)'X'y + y'X(\hat{\beta} - \beta) - \hat{\beta}'X'X\hat{\beta})$$
$$= \frac{1}{Js^2}(0 + 0 + \tilde{\beta}'X'X\tilde{\beta} + 0 + 0 - \hat{\beta}'X'X\hat{\beta})$$
$$= \frac{1}{Js^2}(\tilde{\beta}'X'X\tilde{\beta} - \tilde{\beta}'X'X\hat{\beta} + \tilde{\beta}'X'X\hat{\beta} - \hat{\beta}'X'X\hat{\beta})$$
$$= \frac{1}{Js^2}((\tilde{\beta} - \hat{\beta})'X'X(\tilde{\beta} - \hat{\beta})) = \frac{(R\hat{\beta} - r)'(R(X'X)^{-1}R')^{-1}(R\hat{\beta} - r))/J}{s^2}$$

And hence finished the proof. □

Method: (Test with Large Sample) Under the Assumption about large sample and heteroscedasticity, we use the F estimator to do Hypothesis Test for $H_0 : R\beta - r = 0$, and $H_1 : R\beta - r \neq 0$, where R is a $J \times k$ vector, i.e. reject if $\hat{F} \in [\chi^2_\alpha, +\infty]$, where $\hat{F}$ is defined as:

$$\hat{F}|X = (R\hat{\beta} - r)'(R\hat{V}(\hat{\beta}|X)R')^{-1}(R\hat{\beta} - r)/J|X \to^d \chi^2(J)/J \tag{47}$$

Method: (Test with Large Sample and Homoscedasticity) Under the Assumption about large sample and homoscedasticity, we use the F estimator to do Hypothesis Test for $H_0 : R\beta - r = 0$, and $H_1 : R\beta - r \neq 0$, where R is a $J \times k$ vector, i.e. reject if $\hat{F} \in [\chi^2_\alpha, +\infty]$, where $\hat{F}$ is defined as:

$$\hat{F}|X = \frac{(R\hat{\beta} - r)'(R(X'X)^{-1}R')^{-1}(R\hat{\beta} - r)/J}{s^2}|X \to^d \chi^2(J)/J \tag{48}$$

Claim: (Unbiased and Consistent F-Test) The F-Test described above is unbiased under small sample assumption, and consistent under large sample assumption.

Wald Test for General Non-linear Restriction

Method: (Test with Large Sample) Under the Assumption about large sample and heteroscedasticity, we use the Wald estimator to do Hypothesis Test for $H_0 : g(\beta) = 0$, and $H_1 : g(\beta) \neq 0$, i.e. reject if $\hat{W} \in [\chi^2_\alpha, +\infty]$, where $\hat{W}$ is defined as:

$$\hat{W} = g(\hat{\beta})'(G'\hat{V}(\hat{\beta}|X)G)^{-1}g(\hat{\beta})/J \to^d \chi^2(J)/J \tag{49}$$

, where $G = \partial g(\beta)/\partial\beta|_{\hat{\beta}}$

Method: (Test with Large Sample and Homoscedasticity) Under the Assumption about large sample and homoscedasticity, we use the Wald estimator to do Hypothesis Test for $H_0 : g(\beta) = 0$, and $H_1 : g(\beta) \neq 0$, i.e. reject if $\hat{W} \in [\chi^2_\alpha, +\infty]$, where $\hat{W}$ is defined as:

$$\hat{W} = \frac{g(\hat{\beta})'(G'(X'X)^{-1}G)^{-1}g(\hat{\beta})/J}{s^2} \to^d \chi^2(J)/J \tag{50}$$

, where $G = \partial g(\beta)/\partial\beta|_{\hat{\beta}}$

Claim: (Consistent Wald Test) The Wald Test described above is consistent under large sample assumption.

Restricted Estimation

Restricted Estimation

Theorem: (Restricted Estimation) Suppose the restriction $R\beta = r$ is true, then the restricted regressor is:

$$\tilde{\beta} = \hat{\beta} - (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}(R\hat{\beta} - r) \tag{51}$$

Proof:

The restricted estimator solves the following problem: $min_b \frac{1}{n}(y - Xb)'(y - Xb)\; s.t.\; Rb = r$. Defined the Lagrange function as $L = \frac{1}{2}(y - Xb)'(y - Xb) - \lambda'(Rb - r)$. Take the first order condition we have $X'(y - Xb) - R'\lambda = 0$ and $Rb = r$. Now multiply the first FOC with $R(X'X)^{-1}$, we obtain $R(X'X)^{-1}X'(y - Xb) - R(X'X)^{-1}R'\lambda = 0$, i.e. $R\hat{\beta} = R\tilde{\beta} + R(X'X)^{-1}R'\lambda$, imposing $R\tilde{\beta} = r$ we can solve the Lagrange multiplier $\lambda = (R'(X'X)^{-1}R)^{-1}(R\hat{\beta} - r)$.

Now plug it back into the first FOC, we have $\tilde{\beta} = \hat{\beta} - (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}(R\hat{\beta} - r)$. □

Theorem: (Properties of Restricted Estimation) When the restriction is correct we have:

1. the restricted estimator is consistent
2. $\sqrt{n}(\tilde{\beta} - \beta) \to^d N(0, AQ_{XX}^{-1}\Omega Q_{XX}^{-1}A')$, where $A = I - Q_{XX}^{-1}R'(RQ_{XX}^{-1}R')^{-1}R$.

   Furthermore, if homoscedasticity is true, we have
3. The restricted estimator is more efficient than the OLS Estimator
4. $\sqrt{n}(\tilde{\beta} - \beta) \to^d N(0, \sigma^2 AQ_{XX}^{-1}A')$, where $\sigma^2 AQ_{XX}^{-1}A' = \sigma^2 Q_{XX}^{-1} - \sigma^2 Q_{XX}^{-1}R'(RQ_{XX}^{-1}R')^{-1}RQ_{XX}^{-1} < \sigma^2 Q_{XX}^{-1}$.

Proof:

1. $\tilde{\beta} = \hat{\beta} - (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}(R\hat{\beta} - r)$ when the restriction $R\beta = r$ is true and $\hat{\beta} \to^p \beta$, we have $\tilde{\beta} \to^p \beta$.
2. Note that r does not contribute to the variance of $\tilde{b}eta$, so $\sqrt{n}(\tilde{\beta} - \beta) = \sqrt{n}A\hat{\beta} + \sqrt{n}(X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}r$. So $\sqrt{n}(\tilde{\beta} - \beta) \to^d N(0, AQ_{XX}^{-1}\Omega Q_{XX}^{-1}A')$.
3. the statement 3 is proved by statement 4. □

Note: When the restriction is incorrect the restricted estimator is inconsistent.

Special Case

Definition: (Special Case) For a linear regression model $y = X_1\beta_1 + X_2\beta_2 + e$, suppose we impose the constraint $\beta_2 = 0$, then we have $\tilde{\beta}_1 = (X_1'X_1)^{-1}X_1 y$.

Theorem: (Properties of the Special Case) When the restriction is correct and if homoscedasticity is true, we have

1. the estimator is consistent
2. $\sqrt{n}(\tilde{\beta}_1 - \beta_1) \to^d N(0, \sigma^2 Q_{11}^{-1})$, where $\sigma^2 Q_{11}^{-1} \leq \sigma^2(Q_{11} - Q_{12}Q_{22}^{-1}Q_{21})^{-1}$, i.e. the restricted estimator is more efficient than the original OLS estimator

Proof:

The proof of the property comes from the partitioned regression large sample theory. □

Definition: (Special Case Efficient Estimator) For a linear regression model $y = X_1\beta_1 + X_2\beta_2 + e$, suppose we impose the constraint $\beta_2 = 0$, the most efficient estimator is $\tilde{\beta}_1^* = (X_1'X(X'\Sigma X)^{-1}X'X_1)^{-1}X_1'X(X'\Sigma X)^{-1}X'y$, where $\Sigma = diag(\sigma^2(x_i))$.

Theorem: (Omitted Variables) When the restriction is incorrect, i.e. $\beta_2 \neq 0$, we have

1. Under small sample assumption, the restricted estimator is biased, and $E[\tilde{\beta}_1|X] = \beta_1 + (X_1'X_1)^{-1}(X_1X_2)\beta_2$
2. Under large sample assumption, the restricted estimator is inconsistent, and $lim_p\tilde{\beta}_1|X = \beta_1 + Q_{11}^{-1}Q_{12}\beta_2$

Proof:

We have $\tilde{\beta}_1 = (X_1'X_1)^{-1}X_1 y = (X_1'X_1)^{-1}X_1(X_1\beta_1 + X_2\beta_2 + e)$. Under specific conditions, we can show that the restricted estimator is biased or inconsistent. □

Trinity of Tests

Lagrange Multiplier Test

Definition: (LM Estimator) Define the Lagrange Multiplier Estimator of the restricted estimation as $\tilde{\lambda} = (R(X'X)^{-1}R')^{-1}(R\hat{\beta} - r)$.

Claim: (Properties of LM Estimator) Under large sample assumption, we have:

$$\frac{\tilde{\lambda}}{\sqrt{n}} \to^d N(0, (RQ_{xx}^{-1}R')^{-1}(RQ_{xx}^{-1}\Omega Q_{xx}^{-1}R')(RQ_{xx}^{-1}R')^{-1}) \tag{52}$$

Furthermore, under the Assumption about large sample and homoscedasticity, we have:

$$\frac{\tilde{\lambda}}{\sqrt{n}} \to^d N(0, \sigma^2(R(X'X)^{-1}R')^{-1}) \tag{53}$$

Method: (Lagrange Multiplier Test) Under the Assumption about large sample and heteroscedasticity, we use the LM statistic to do Hypothesis Test for $H_0 : R\beta - r = 0$, and $H_1 : R\beta - r \neq 0$, where R is a $J \times k$ vector, i.e. reject if $L\hat{M} \in [\chi^2_\alpha, +\infty]$, where $L\hat{M}$ is defined as:

$$L\hat{M} = \frac{\tilde{\lambda}'\hat{V}_\lambda^{-1}\tilde{\lambda}/J}{n} \to^d \chi^2(J)/J \tag{54}$$

where $\hat{V}_\lambda = (R(X'X)^{-1}R')^{-1}(R\hat{V}(\tilde{\beta}|X)R')(R(X'X)^{-1}R')^{-1}/n$ is the variance estimator of the restricted regression.

Method: (Lagrange Multiplier Test with Homoscedasticity) Under the Assumption about large sample and homoscedasticity, we use the LM statistic to do Hypothesis Test for $H_0 : R\beta - r = 0$, and $H_1 : R\beta - r \neq 0$, where R is a $J \times k$ vector, i.e. reject if $L\hat{M} \in [\chi^2_\alpha, +\infty]$, where $L\hat{M}$ is defined as:

$$L\hat{M} = \frac{\tilde{\lambda}'\hat{V}_\lambda^{-1}\tilde{\lambda}/J}{n} = \frac{(R\hat{\beta} - r)'(R(X'X)^{-1}R')^{-1}(R\hat{\beta} - r)/J}{\tilde{s}^2} \to^d \chi^2(J)/J \tag{55}$$

where $\tilde{s}^2 = SSE_R/n - (k - J)$ is the variance of the residual of the restricted estimation.

Likelihood Ratio Test

Definition: (LR Estimator) Under homoscedasticity and gaussian error assumption, define the Likelihood Ratio Estimator of the restricted estimation as $LR = 2(lnL(\hat{\beta}, \hat{\sigma}^2) - lnL(\tilde{\beta}, \tilde{\sigma}^2))$. note that here $(\hat{\beta}, \hat{\sigma}^2)$ is the MLE estimator.

Theorem: (LR Estimator and F Statistic) We have $LR = nlog(1 + JF/(n - k))$.

Proof:

Note that $lnL(\hat{\beta}, \hat{\sigma}^2) = -\frac{n}{2}(ln(SSE_U/n) + ln(2\pi) + 1)$ and $lnL(\tilde{\beta}, \tilde{\sigma}^2) = -\frac{n}{2}(ln(SSE_R/n) + ln(2\pi) + 1)$. So we can plug them in and get $LR = 2(lnL(\hat{\beta}, \hat{\sigma}^2) - lnL(\tilde{\beta}, \tilde{\sigma}^2)) = nln(SSE_R/SSE_U) = nln(1 - \frac{J}{n-k}\frac{(SSE_R - SSE_U)/J}{SSE_U/(n-k)})$. Hence we have $LR = nlog(1 + JF/(n - k))$.

By Taylor expansion of a log function, we have $LR \approx n/(n - k)F$. $\square$

Method: (Likelihood Ratio Test with Homoscedasticity and Gaussian Error) Under the Assumption about large sample and homoscedasticity and Gaussian Error, we use the LR statistic to do Hypothesis Test for $H_0 : R\beta - r = 0$, and $H_1 : R\beta - r \neq 0$, where R is a $J \times k$ vector, i.e. reject if $L\hat{R} \in [\chi^2_\alpha, +\infty]$, where $L\hat{M}$ is defined as:

$$L\hat{R} = \frac{(R\hat{\beta} - r)'(R(X'X)^{-1}R')^{-1}(R\hat{\beta} - r)/J}{\hat{\sigma}^2} \to^d \chi^2(J)/J \tag{56}$$

where $\hat{\sigma}^2$ is the variance of the MLE of σ^2 under the unrestricted estimation.

Note: As n increases, s^2, $\tilde{s}^2$ and $\hat{\sigma}^2$ are all consistent estimator of σ^2. Hence Wald Test, LM Test and LR Test are all consistent and are similar to each other.

Confidence Interval

Definition: (Confidence Interval) Given the data $\{S_n\}$ we observe, suppose $S_i \sim f(\theta)$. Let L and U be two statistics. We say (L, U) is a $1-\alpha$ Confidence Interval for $g(\theta)$ if $P(g(\theta) \in (L, U)) = 1-\alpha$.

Special Issues in OLS

Functional Form

Non-linearities

Method: (High Order Regression) Suppose a model is $y = \beta_0 + x\beta_1 + x^2\beta_2 + x^3\beta_3 + \epsilon$. One can use the OLS estimator to estimate this equation since it is still linear in parameter.

Method: (Interaction) Suppose a model is $y = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_1x_2\beta_3 + \epsilon$. One can use the OLS estimator to estimate this equation since it is still linear in parameter.

Method: (Dummy Variables) Suppose a model is $y = \beta_0 + x_1\beta_1 + \epsilon$, where x_1 is a dummy variable. One can use the OLS estimator to estimate this equation since it is still linear in parameter.

Method: (Category Variables) Suppose a model is $y = \beta_0 + x_1\beta_1 + \epsilon$, where x_1 is a category variable with $x_1 = 0, 1, 2, \ldots, k$. One can use the OLS estimator to estimate $y = \beta_0 + x_1 1\beta_1 + x_1 2\beta_2 + \ldots + x_1 k\beta_k + \epsilon$.

Note: Remember to leave one category out.

Difference in Difference

Method: (Difference in Difference) Suppose a model is $y = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_1x_2\beta_3 + \epsilon$. x_1, x_2 are two dummy variables, the first is the policy dummy and the second is the trend dummy. Assuming the trend effect is parallel, we can estimate the effect of x_1 with β_3.

Testing for Functional Form

Method: (Ramsey RESET Test) To test if the functional form is correct, we first run the OLS with $y = x\beta + \epsilon$. Next, get the predictor $\hat{y} = x\hat{\beta}$. Then regress $y = x\gamma_1 + \hat{y}^2\gamma_2 + \hat{y}^3\gamma_3 + \hat{y}^4\gamma_4 + \mu$ and do a F test on $H_0 : \gamma_2 = 0, \gamma_3 = 0, \gamma_4 = 0$.

Bootstrapping

Method: (Bootstrapping)

1. From the original sample $\{X_1, \ldots, X_n\}$ generate an estimator $\hat{\theta} = h(X_1, \ldots X_n)$
2. Take a random sample of the same size n from the original sample with replacement, and form a new sample $\{X_1^1, \ldots, X_n^1\}$, get an estimator $\hat{\theta}^1 = h(X_1^1, \ldots X_n^1)$
3. Repeat step 2 and form a new sample $\{X_1^k, \ldots, X_n^k\}$, get estimators $\hat{\theta}^k = h(X_1^k, \ldots X_n^k)$
4. Compute the distribution with the estimators θ^k
5. Use the distribution calculated above to do Hypothesis Test or give the Confidence Interval

Claim: (Bootstrapping Theory) When the time bootstrapping repeats increases, the bootstrapping distribution converges to the distribution of the real estimator.

Efficient Estimator with Heteroskedasticity

Method: (Testing for Heteroskedasticity) Consider a model $y = x\beta + \epsilon$, first do the OLS regression as usual. Then get the predicted residual $\hat{e}_i$. Now regress $\hat{e}^2 = \gamma_0 + x\gamma_1 + \mu$. Now test for heteroskedasticity, with $H_0 : E[e^2|X] = \sigma^2$, by doing a F test on $\gamma_1 = 0$.

Method: (WLS Estimator) Suppose heteroskedasticity is true, then

1. Do OLS of y on x and get the estimated residual $\hat{e}$
2. Create $ln(\hat{e}^2)$ and run OLS of $ln(\hat{e}^2)$ on x to get fitted value $\hat{g}$
3. Estimate σ_i^2 with $\hat{\sigma}_i^2 = e^{\hat{g}_i}$
4. Do WLS using the estimated weight in the last step

Further Issues

Predictions

Claim: (Prediction) The forecast estimator for a single data point is $\hat{y}_i = x_i\hat{\beta}$. We have:

1. $AVar(\hat{y}_i - y_i|X) = x_i AVar(\hat{\beta})x_i' + Var(e_i|x_i)$
2. Under homoscedasticity, we have $AVar(\hat{y}_i - y_i|X) = x_i AVar(\hat{\beta})x_i' + \sigma^2$

Clustering

Definition: (Clustering Issue) When the i.i.d. assumption is violated it is called to have a Clustering Issue.

Note: Heteroskedasticity is a special case for clustering issue. The correlation between two observations can be not zero.

Multicollinearity

Claim: (Multicollinearity) Consider the partitioned model $y = x_1'\beta_1 + x_K\beta_K + \epsilon$, assuming homoscedasticity, we have $Var(\hat{\beta}_K|X) = \sigma^2/((1-R_K^2)x_K'M_0x_K)$, where $R_K^2 = 1-(x_K'M_1x_K)/(x_K'M_0x_K)$ is the R squared regressing x_K on x_1.

Note: This implies that when one of the independent variable X can be predicted pretty well by other independent variables, the variance of the estimator $\hat{\beta}$ would be high. So the estimation might be less precise.

Endogeneity

Source of Endogeneity

Omitted Variables

Definition: (Omitted Variables) For a linear regression model $y = X_1\beta_1 + X_2\beta_2 + e$, suppose we omitted X_2 from the OLS. The OLS Estimator is called having Omitted Variable issue.

Theorem: (Omitted Variables Issues) Suppose we have $X_2 = X_1\delta + \mu$, the OLS estimator have the following properties:

1. Under small sample assumption, the estimator is biased, and $E[\hat{\beta}_1|X_1] = \beta_1 + \delta\beta_2$
2. Under large sample assumption, the estimator is inconsistent, and $lim_p\hat{\beta}_1|X_1 = \beta_1 + \delta\beta_2$

Proof:

We have $\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'y = (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + e) = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'e$. We have $E[e|X] = 0$ and $X'e|X \to^p 0$, hence we have what we want to show. □

Errors in Variables

Definition: (Errors in Variables) For a linear regression model $y = X\beta + e$, suppose we can only observe a noisy signal S of X. The OLS Estimator of regression Y on S is called having Errors in Variables issue.

Theorem: (Errors in Variables Issues) Suppose we have that $S = X + u$, then under large sample assumption, the OLS estimator is inconsistent, and $lim_p \hat{\beta} = \beta\frac{\sigma_X^2}{\sigma_X^2+\sigma_u^2}$.

Proof:

We have $\hat{\beta} = (S'S)^{-1}S'y = ((X+u)'(X+u))^{-1}(X+u)'(X\beta+e)$. So $(S'S/n)^{-1} \to^p (\sigma_X^2 + \sigma_u^2)^{-1}$, and $S'e \to^p 0$, and $S'X\beta \to^p \sigma_X^2\beta$. Combine them we have $lim_p \hat{\beta} = \beta\frac{\sigma_X^2}{\sigma_X^2+\sigma_u^2}$. □

Simultaneity

Definition: (Simultaneity) For a linear regression system $Q = P\beta_1 + e_1$ and $Q = P\beta_2 + e_2$, suppose we omitted X_2 from the OLS. The OLS Estimator is called having simultaneity issue.

Note: When we have Simultaneity issues, we cannot run OLS.

Instrument Variable

Definition: (Instrument Variable) Consider a linear regression model $Y = X_1\beta_1 + X_2\beta_2 + e$, with $E[e|X_2] \neq 0$ and β_2 is $k \times 1$. Suppose we have another set of data Z, which is $J \times 1$ and $J \geq k$, and we have $E[Ze] = 0$, but $E[ZX_{2k}] \neq 0$, then Z is called an Instrument Variable. Furthermore, suppose we have $X_2 = X_1\Gamma_1 + Z\Gamma_2 + u$ then we have the following linear regression $Y = X_1(\beta_1 + \Gamma_1\beta_2) + Z\Gamma_2\beta_2 + (e + \beta_2 u) = X_1\gamma_1 + Z\gamma_2 + v$.

Definition: (Identification) Let Γ_2 be a $J \times k$ metrics. Suppose the following conditions are satisfied:

1. Order Condition: $J \geq k$
2. Rank Condition: $rank(\Gamma_2) = k$

We say the endogenous variable X_2 is identified.

Definition: (IV Estimator) When the endogenous variable is identified, we can define the IV Estimator as:

1. if $J = k$, define $\hat{\beta}_2^{IV} = \hat{\Gamma}_2^{-1}\hat{\gamma}_2$
2. if $J > k$, define $\hat{\beta}_2^{IV} = (\hat{\Gamma}_2'A\hat{\Gamma}_2)^{-1}\hat{\Gamma}_2'A\gamma_2$, where A is a symmetric and positive definite

General Method of Moments

GMM Estimator

Definition: (General Method of Moments) Suppose we have $\frac{1}{n}\sum_{i=1}^n z_i(y_i - x_i'\beta) = 0$. Let W_n be a symmetric positive definite matrix, General Method of Moments estimator is defined as:

$$\bar{\beta} = argmin_\beta\{(y - X\beta)'ZW_nZ'(y - X\beta)\} \tag{57}$$

Note: We only derive the GMM Estimator under the large sample assumptions.

Assumption: (Large Sample Assumption of General Method of Moments)

1. $W_n \to W$ and W is symmetric and positive definite
2. $E[z_i e_i] = 0$
3. $E[z_i x_i'] = Q_{zx}$ exists
4. $E[z_i z_i' e_i^2] = \Omega < +\infty$

Theorem: (GMM Estimator) Under the Assumption of General Method of Moments, the GMM estimator is

$$\bar{\beta} = (X'ZW_nZ'X)^{-1}X'ZW_nZ'y \tag{58}$$

Proof:

GMM Estimator solves $min_\beta\{(y - X\beta)'ZW_nZ'(y - X\beta)\}$. The first order condition is $X'ZW_nZ'(y - X\beta) = 0$ which will give us what we want to show. □

Theorem: (GMM Estimator Property) Under the Large Sample Assumption of General Method of Moments, we have

1. $\sqrt{n}(\bar{\beta} - \beta) \to^d N(0, (Q_{zx}'WQ_{zx})^{-1}Q_{zx}'W\Omega WQ_{zx}(Q_{zx}'WQ_{zx})^{-1})$
2. $lim_p\hat{Q}_{zx} = lim_p Z'X/n = Q_{zx}$
3. $lim_p\hat{\Omega} = lim_p \sum_{i=1}^n \hat{e}_i^2 z_i z_i'/n = \Omega$, where $\hat{e} = y - X\bar{\beta}$

Proof:

1. $\sqrt{n}(\bar{\beta} - \beta) = \sqrt{n}((X'ZW_nZ'X)^{-1}X'ZW_nZ'e)$. Note that $(X'ZW_nZ'X/n^2)^{-1} \to^p (Q_{zx}'WQ_{zx})^{-1}$, and $(X'Z/n)W_n \to^p Q_{zx}'W$, and $\sqrt{n}(Z'e/n) \to^d N(0, \Omega)$. Combine them we get what we want to show.
2. This is true by law of large number.
3. This is true by law of large number. □

Note: From 2 and 3 generate a consistent estimator of the asymptotic variance of the estimator $\bar{\beta}$.

Special Case

Claim: (Special Case) Under the Assumption of General Method of Moments, if $J = K$, the GMM estimator is

$$\bar{\beta} = (Z'X)^{-1}Z'Y \tag{59}$$

Theorem: (Special Case Property) Under the Large Sample Assumption of General Method of Moments, if $J = K$, we have

1. $\sqrt{n}(\bar{\beta} - \beta) \to^d N(0, (Q_{zx}'\Omega^{-1}Q_{zx})^{-1})$
2. $lim_p\hat{Q}_{zx} = lim_p Z'X/n = Q_{zx}$
3. $lim_p\hat{\Omega} = lim_p \sum_{i=1}^n \hat{e}_i^2 z_i z_i'/n = \Omega$, where $\hat{e} = Y - X\bar{\beta}$

Proof:

Just apply the properties of GMM under the special case. □

Note: From 2 and 3 generate a consistent estimator of the asymptotic variance of the estimator $\bar{\beta}$.

Efficient GMM Estimator

Theorem: (Optimal Weight Matrix) We have that for any W,

$$(Q_{zx}'WQ_{zx})^{-1}Q_{zx}'W\Omega WQ_{zx}(Q_{zx}'WQ_{zx})^{-1} \geq (Q_{zx}'\Omega^{-1}Q_{zx})^{-1} \tag{60}$$

Proof:

We want to show $(Q_{zx}'WQ_{zx})(Q_{zx}'W\Omega WQ_{zx})^{-1}(Q_{zx}'WQ_{zx}) \leq Q_{zx}'\Omega^{-1}Q_{zx}$. We can show that:

$$\begin{aligned} &Q_{zx}'\Omega^{-1}Q_{zx} - (Q_{zx}'WQ_{zx})(Q_{zx}'W\Omega WQ_{zx})^{-1}(Q_{zx}'WQ_{zx}) \\ &= Q_{zx}'\Omega^{-\frac{1}{2}}(I - \Omega^{-\frac{1}{2}}Q_{zx}(Q_{zx}'W\Omega WQ_{zx})^{-1}Q_{zx}'\Omega^{-\frac{1}{2}})\Omega^{-\frac{1}{2}}Q_{zx} \\ &= A'(I - B(B'B)^{-1}B')A = A'M_BA' \geq 0 \end{aligned} \tag{61}$$

because $A'M_BA'$ is the SSE of some regression, and SSE are positive semi-definite. $\square$

Definition: (Feasible Efficient GMM Estimator) The feasible efficient estimator is $\bar{\beta} = (X'Z\hat{\Omega}^{-1}Z'X)^{-1}X'Z\hat{\Omega}^{-1}Z'Y$.

Theorem: (Efficient GMM Estimator Property) Under the Large Sample Assumption of GMM Estimator we have

1. $\sqrt{n}(\bar{\beta} - \beta) \to^d N(0, (Q'_{zx}\Omega^{-1}Q_{zx})^{-1})$
2. $lim_p \hat{Q}_{zx} = lim_p Z'X/n = Q_{zx}$
3. $lim_p \hat{\Omega} = lim_p \sum_{i=1}^n \hat{e}_i^2 z_i z_i'/n = \Omega$, where $\hat{e} = Y - X\bar{\beta}$

Proof:

Just apply the properties of GMM to $W_n = \Omega^{-1}$. $\square$

Note: From 2 and 3 generate a consistent estimator of the asymptotic variance of the estimator $\bar{\beta}$.

Note: The Optimal Weight Matrix is such that $W_n \to \Omega^{-1}$.

2SLS Estimator

Definition: (2SLS Estimator) the 2 Stage Least Square Estimator is defined as

$$\hat{\beta}^{2SLS} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y \tag{62}$$

Theorem: (2SLS and GMM) 2SLS Estimator is GMM Estimator with $W_n = (Z'Z/n)^{-1}$, which is optimal if Homoscedasticity is true, i.e. $E[z_i z_i' e_i^2] = \sigma^2 E[z_i z_i']$.

Proof:

The 2SLS estimator is defined with 2 stages. First regress X on Z, we have $\hat{X} = P_Z X$. Then regress y on $\hat{X}$, we will then get the 2 stage least square estimator, i.e. $\hat{\beta} = (\hat{X}'\hat{X})^{-1}\hat{X}'y = (X'P_Z X)^{-1}X'P_Z y$. $\square$

Identification Issues

Weak IV

Definition: (Weak Identification) When the rank condition is not satisfied, i.e. $rank(\Gamma_2) < k$, we say that the IVs are weak.

Theorem: (Weak IV Problem) When $X_1 = 0,\ \Gamma_2 = \delta/\sqrt{n} \to 0$, the GMM Estimator is inconsistent.

Proof:

For simplicity we prove it with the special case when $J = K$ and $X = X_2$. We have $\bar{\beta} = (Z'X)^{-1}Z'y$ and $\bar{\beta} - \beta = (Z'X)^{-1}Z'e = (\Gamma Z'Z + Z'u)^{-1}Z'e$. Then $\delta Z'Z/n \to^p \delta E[z_i^2] \neq 0$, $\sqrt{n}Z'u/n \to^d N(0, E[z_i^2 u_i^2])$ and $\sqrt{n}Z'e/n \to^d N(0, E[z_i^2 e_i^2])$. Combine them we can conclude that $\bar{\beta} \nrightarrow^p \beta$. $\square$

Definition: (Weak IV Test) To test if the IVs are weak, we can take the regression $X_2 = X_1\Gamma_1 + Z\Gamma_2 + u$, and do a Wald Test or F test with $H_0 : \Gamma_2 = 0$.

Hansen's J Test

Definition: (Over Identification) When we have more IVs than the endogenous variables, i.e. $J > k$, we say that the endogenous variables are over identified.

Definition: (Hansen's J) Define Hansen's J statistic as $J = n(y - X\bar{\beta})'Z\hat{\Omega}^{-1}Z'X\hat{\Omega}^{-1}X'Z\hat{\Omega}^{-1}Z'(y - X\beta)$.

Theorem: (Hansen's J Property) Under the large sample assumption of General Method of Moments, we have $J \to^d \chi^2(J - k)$.

Proof:

For simplicity we add homoscedasticity and try to prove this with the 2SLS estimator.

Under homoscedasticity the statement Hansen's J statistic is defined as $J = (e - \bar{e})' P_Z (e - \bar{e})$, where $\bar{e} = X(X'P_Z X)^{-1} X' P_Z e$. So we have $J = e'Z(Z'Z)^{-\frac{1}{2}}(I - (Z'Z)^{\frac{1}{2}} Z'X(X'Z(Z'Z)^{-\frac{1}{2}}(Z'Z)^{-\frac{1}{2}}Z'X)^{-1}X'Z(Z'Z)^{-\frac{1}{2}})(Z'Z)^{-\frac{1}{2}}Z'e = e'B_n'(I - B_n(B_n'B_n)^{-1}B_n)B_n e$. Note that we have $B_n \to^p B = Q_{ZZ}^{-\frac{1}{2}} Q_{ZX}$. Note that this implies $(I - B_n(B_n'B_n)^{-1}B_n) \to^p M_B$. Since M_B is symmetric and idempotent, we can write $M_B = H\Lambda H'$ where $H'H = I$ and

$$\Lambda = \begin{pmatrix} I_{n-k} & 0 \\ 0 & 0 \end{pmatrix} \tag{63}$$

Since $Trace(M_B) = Trace(HH'\Lambda) = Trace(\Lambda) = n - k$.

Now since $Z'e/\sqrt{n} \to^d N(0, \sigma^2 Q_{ZZ})$ and $(Z'Z/n)^{-\frac{1}{2}} \to^p Q_{ZZ}^{-\frac{1}{2}}$. Combine everything together we have $J \to^d \chi^2(J-k)$. $\square$

Definition: (Hansen's J Test) Under the large sample assumption of General Method of Moments, we use the J estimator to do Hypothesis Test for $H_0 : E[z_i e_i] = 0$, and $H_1 : E[z_i e_i] \neq 0$, i.e. reject if $\hat{J} \in [\chi^2_\alpha, +\infty]$, where $\hat{J}$ is defined as:

$$\hat{J} = n(Y - X\bar{\beta})' Z \hat{\Omega}^{-1} Z'(Y - X\bar{\beta}) \to^d \chi^2(J-k) \tag{64}$$

Hausman Test

Definition: (Hausman Test) Under the large sample assumption of General Method of Moments, we do a Hypothesis Test for $H_0 : \hat{\beta} = \bar{\beta}$, and $H_1 : \hat{\beta} \neq \bar{\beta}$, i.e. if there are endogeneity or not, we define a Hausman statistic:

$$H = n(\hat{\beta} - \bar{\beta})' V^+ (\hat{\beta} - \bar{\beta}) \to^d \chi^2(k) \tag{65}$$

where $V^+ = V(\hat{\beta} - \bar{\beta})^+ = (V(\hat{\beta}) - V(\bar{\beta}))^+$ is the G-inverse of V.

Definition: (Alternative Hausman Test) Under the large sample assumption of General Method of Moments, we do a Hypothesis Test for $H_0 : E[z_i e_i] = 0$, and $H_1 : E[x_i e_i] \neq 0$, i.e. if there are endogeneity or not, by doing OLS on $y = X_1\beta_1 + X_2\beta_2 + \hat{u}\rho + \epsilon$, where $\hat{u}$ is the OLS residual from regressing $X_2 = X_1\Gamma_1 + Z\Gamma_2 + u$, and do a F Test with ρ.

Claim: (Relationship Between Alternative Hausman Test and 2SLS) If we do the regression of $y = X_1\beta_1 + X_2\beta_2 + \hat{u}\rho + \epsilon$, the estimator $\hat{\beta}$ will be the 2SLS estimator.